Determining the Biases and Consistencies in the Evidence for Conservation

UNIVERSITY OF CAMBRIDGE

Alec Philip Christie

King's College

January 2021

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Determining the Biases and Consistencies in the Evidence for Conservation

Alec Philip Christie

Summary

Earth's biodiversity is facing an anthropogenic extinction crisis and yet conservation efforts are chronically underfunded. There is a tremendous need to act as efficiently as possible to conserve biodiversity – evidence-based approaches are essential to this mission. Biodiversity conservation is undergoing an evidence-based revolution, emulating techniques to synthesise evidence pioneered in medicine, such as systematic reviews, meta-analyses, and subject-wide evidence syntheses, which have had success in summarising the evidence on what works in conservation. To date, however, the biases and consistencies in this evidence base have neither been quantified nor explored in detail. This is important to facilitate further evidence-based decision-making in conservation and to improve the reliability and relevance of the evidence base. In line with its title, this thesis is structured into quantifying and addressing two types of biases in the evidence base for conservation: within-study and between-study biases. Both types of biases represent fundamental challenges that must be overcome to ensure evidence-based decision-making becomes more commonplace in biodiversity conservation practice and policy.

Within-study biases affect the reliability (internal validity) of research findings, which are known to hinge upon the choice of study design used to collect data. Many study designs are used to test the effectiveness of conservation interventions, including 'gold standard' randomised experiments (in the medical sciences) and various types of observational designs (often used when randomised experiments are too hard to implement cheaply, ethically, or practically). However, no large-scale, direct, and quantitative comparisons of the relative reliability of different study designs have been made and therefore little is known about how much more trust should be placed in results obtained using one design over another. I tackle this issue by quantitatively estimating the relative reliability of results obtained by commonly used study designs in ecology. In the first Chapter, I simulate the performance of different study designs using empirically derived estimates of the magnitude of study design bias from 51 ecological datasets obtained from a range of studies around the world. In the second Chapter, I build on these simulations by digging deeper into the raw datasets, conducting pairwise comparisons of the estimates given by different designs within each dataset. I also develop a hierarchical Bayesian model to quantify the relative reliability of study designs, enabling meta-analyses to

account more effectively for the bias and variance introduced by studies. This approach attempts to tackle the challenging issue of combining study results obtained using different study designs, which has been a hotly debated issue in evidence synthesis.

Understanding between-study biases, or biases affecting the wider literature's distribution and coverage, is crucial to prioritising future conservation research and action. In my third Chapter, I use the Conservation Evidence database (comprised of quantitative tests of conservation interventions) to quantify the spatial, taxonomic, bioclimatic, and design-related biases to show the severity of the knowledge gaps for amphibian and bird conservation. Entire orders of amphibians and birds were either poorly represented or absent in the evidence base, whilst more credibly designed studies were located, almost exclusively, in North America, Europe, and Australasia. In addition, fewer studies were conducted in locations with more threatened amphibian and bird species. These results run counter to the mission of conservation, suggesting that places with the greatest need for conservation often lack credible evidence. In my fourth Chapter, I investigate how much evidence exists for certain local questions. This is important because decision-makers typically prefer evidence that is locally valid and relevant to their specific setting. I quantify how much evidence exists within certain distances of a given decision-maker anywhere in the world, and then demonstrate, on average, how little relevant or credible evidence exists for most decision-makers. This work reinforces that there is a serious mismatch between where we test conservation interventions and where they are needed, and that providing decision-makers with locally relevant evidence is a major challenge.

This thesis demonstrates the fundamental importance of study design in determining the reliability of study findings, whilst also highlighting important knowledge gaps and biases in the literature that tests conservation interventions. Based on the findings of this thesis, I provide several recommendations and possible solutions to improve the evidence base for conservation, and to ensure that evidence-based decision-making and practice becomes more widespread and successful.

Table of Contents

Preface

As Chapters 2-5 were prepared for publication and involved collaborations, I use the pronoun ‘we’ rather than ‘I’ throughout. Supplementary Information for each Chapter are located at the end of each Chapter. Figures, Tables, and Appendices with an ‘S’ preceding the number (e.g., Figure S1, Table S1, Appendix S1) refer to the Supplementary Information within that Chapter. Chapter 1 is the Introductory Chapter, Chapters 2-5 are Data Chapters, and Chapter 6 is the Concluding Chapter.

Statement of contributions

I conceived, designed, and conducted the research behind each Chapter and led the writing of this thesis at all stages. The roles of people other than myself in this work are described as follows.

Entire thesis

William Sutherland, Philip Martin, and Tatsuya Amano contributed to the conception, planning, and design of all research. William Sutherland commented on Chapters 1-6, whilst Philip Martin and Tatsuya Amano commented on Chapters 2-5. Further details of contributions to individual Chapters are outlined below.

Chapter 2

Gorm Shackelford and Benno Simmons contributed to the design of analyses and commented on drafts of the paper that is the basis of this Chapter. Environmental datasets used in this Chapter’s analyses were collected and contributed by: P. Edwards, G.R. Hodgson, H. Welsh, J.V. Vieira, M. van Deurs, T.M. Grome, M. Kaspersen, H. Jensen, C. Stenberg, T.K. Sørensen, J. Støttrup, T. Warnar, H. Mosegaard, A. Schwerk, A. Velando, Dolores River Restoration Partnership, J.S. Pinilla, A. Page, M. Dasey, D. Maguire, J. Barlow, J. Louzada, Jari Florestal, R.T. Buxton, C.R. Schacter, J. Seoane, M.G. Conners, K. Nickel, G. Marakovich, A. Wright, G. Soprone, CSIRO, A. Elosegi, L. García-Arberas, J. Díez, A. Rallo, Parks and Wildlife Finland, Parc Marin de la Côte Bleue, D. Abecasis, M. Adjeroud, J.C. Alonso, A. Anton, B.P. Baldigo, R. Barrientos, J.E. Bicknell, D.A. Buhl, J. Cebrian, R.S. Ceia, L. Cibils-Martina, S. Clarke, J. Claudet, M.D. Craig, D. Davoult, A. De Backer, M.K. Donovan, T.D. Eddy, F.M. França, J.P.A. Gardner, B.P. Harris, A. Huusko, I.L. Jones, B.P. Kelaher, J.S. Kotiaho, A. López-Baucells, H.L. Major, A. Mäki-Petäys, B. Martín, C.A. Martín, D. Mateos-Molina, R.A. McConnaughey, M. Meroni, C.F.J. Meyer, K. Mills, M. Montefalcone, N. Noreika, C. Palacín, A. Pande, C.R. Pitcher, C. Ponce, M. Rinella, R. Rocha, M.C. Ruiz-Delgado, J.J. Schmitter-Soto, J.A. Shaffer, S. Sharma, A.A. Sher, D. Stagnol, T.R. Stanley, K.D.E. Stokesbury, A. Torres, O. Tully, T. Vehanen, and C. Watts.

Chapter 3

Qingyuan Zhao, Statistics Laboratory, University of Cambridge, assisted with the development of analyses and writing code for hierarchical Bayesian modelling.

The past and present members of the Conservation Evidence project (www.conservationevidence.com) collected metadata from studies testing conservation interventions and summarised their study design, which I used to assess the prevalence of different study designs.

Environmental datasets for the analyses presented in this Chapter were collected and contributed by the following organisations and people, many of whom also commented on drafts of the paper that is the basis of this Chapter: P. Edwards, G.R. Hodgson, H. Welsh, J.V. Vieira, M. van Deurs, T.M. Grome, M. Kaspersen, H. Jensen, C. Stenberg, T.K. Sørensen, J. Støttrup, T. Warnar, H. Mosegaard, A. Schwerk, A. Velando, Dolores River Restoration Partnership, J.S. Pinilla, A. Page, M. Dasey, D. Maguire, J. Barlow, J. Louzada, Jari Florestal, R.T. Buxton, C.R. Schacter, J. Seoane, M.G. Conners, K. Nickel, G. Marakovich, A. Wright, G. Soprone, CSIRO, A. Elosegi, L. García-Arberas, J. Díez, A. Rallo, Parks and Wildlife Finland, Parc Marin de la Côte Bleue, D. Abecasis, M. Adjeroud, J.C. Alonso, A. Anton, B.P. Baldigo, R. Barrientos, J.E. Bicknell, D.A. Buhl, J. Cebrian, R.S. Ceia, L. Cibils-Martina, S. Clarke, J. Claudet, M.D. Craig, D. Davoult, A. De Backer, M.K. Donovan, T.D. Eddy, F.M. França, J.P.A. Gardner, B.P. Harris, A. Huusko, I.L. Jones, B.P. Kelaher, J.S. Kotiaho, A. López-Baucells, H.L. Major, A. Mäki-Petäys, B. Martín, C.A. Martín, D. Mateos-Molina, R.A. McConnaughey, M. Meroni, C.F.J. Meyer, K. Mills, M. Montefalcone, N. Noreika, C. Palacín, A. Pande, C.R. Pitcher, C. Ponce, M. Rinella, R. Rocha, M.C. Ruiz-Delgado, J.J. Schmitter-Soto, J.A. Shaffer, S. Sharma, A.A. Sher, D. Stagnol, T.R. Stanley, K.D.E. Stokesbury, A. Torres, O. Tully, T. Vehanen, and C. Watts.

Chapter 4 and Chapter 5

Silviu Petrovan, Gorm Shackelford, Benno Simmons, Rebecca Smith, David Williams, and Claire Wordley commented on all drafts of these Chapters. Benno Simmons and David Williams also contributed to the design of the analyses presented. Anne-Christine Mupepele provided useful comments on initial drafts of the papers that are the basis of these Chapters. Past and present members of the Conservation Evidence project collected metadata from studies testing conservation interventions which I analysed in this Chapter.

Publications

The following publications form the basis of this thesis:

Chapter 2: Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Chapter 3: Christie, A.P., Abecasis, D., Adjeroud, M., Alonso, J.C., Amano, T., Anton, A., Baldigo, B.P., Barrientos, R., Bicknell, J.E., Buhl, D.A., Cebrian, J., Ceia, R.S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M.D., Davoult, D., De Backer, A., Donovan, M.K., Eddy, T.D., França, F.M., Gardner, J.P.A., Harris, B.P., Huusko, A., Jones, I.L., Kelaher, B.P., Kotiaho, J.S., López-Baucells, A., Major, H.L., Mäki-Petäys, A., Martín, B., Martín, C.A., Martin, P.A., Mateos-Molina, D., McConnaughey, R.A., Meroni, M., Meyer, C.F.J., Mills, K., Montefalcone, M., Noreika, N., Palacín, C., Pande, A., Pitcher, C.R., Ponce, C., Rinella, M., Rocha, R., Ruiz-Delgado, M.C., Schmitter-Soto, J.J., Shaffer, J.A., Sharma, S., Sher, A.A., Stagnol, D., Stanley, T.R., Stokesbury, K.D.E., Torres, A., Tully, O., Vehanen, T., Watts, C., Zhao, Q., Sutherland, W.J., 2020. Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. Nature Communications 11, 6377. https://doi.org/10.1038/s41467-020-20142-y

Chapter 4: Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020. The challenge of biased evidence in conservation. Conservation Biology cobi.13577. https://doi.org/10.1111/cobi.13577

Chapter 5: Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020. Poor availability of context-specific evidence hampers decision-making in conservation. Biological Conservation 248, 108666. https://doi.org/10.1016/j.biocon.2020.108666

In addition, I have contributed to the following manuscripts and publications during my PhD:

Christie, A.P.,* White, T.B.,* et al. *Pre-print*. Reducing publication delay to improve the efficiency and impact of conservation science. BioRxiv. https://doi.org/10.1101/2021.03.30.437223 *Joint first authors.

Amano, T., Espinola, V.B., **Christie, A.P.**, et al. *In review.* The neglected role of non-English-language science in conserving global biodiversity.

Christie, A.P. et al. *In prep.* A practical conservation tool to combine diverse forms of evidence for rapid, systematic, and transparent decisions.

Shackelford, G.E., Martin, P.A., Hood, A.S.C., **Christie, A.P.**, Kulinskaya, E., Sutherland, W.J. 2021. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biology 19, 33. https://doi.org/10.1186/s12915-021-00974-w

Downey, H., Amano, T., Cadotte, M., Cook, C.N., Cooke, S.J., Haddaway, N.R., Jones, J.P.G., Littlewood, N., Walsh, J.C., Abrahams, M.I., Adum, G., Akasaka, M., Alves, J.A., Antwis, R.E., Arellano, E.C., Axmacher, J., Barclay, H., Batty, L., Benítez-López, A., Bennett, J.R., Berg, M.J., Bertolino, S., Biggs, D., Bolam, F.C., Bray, T., Brook, B.W., Bull, J.W., Burivalova, Z., Cabeza, M., Chauvenet, A.L.M., **Christie, A.P.**, Cole, L., Cotton, A.J., Cotton, S., Cousins, S.A.O., Craven, D., Cresswell, W., Cusack, J.J., Dalrymple, S., Davies, Z.G., Diaz, A., Dodd, J.A., Felton, A., Fleishman, E., Gardner, C.J., Garside, R., Ghoddousi, A., Gilroy, J.J., Gill, D.A., Gill, J.A., Glew, L., Grainger, M.J., Grass, A.A., Greshon, S., Gundry, J., Hart, T., Hopkins, C.R., Howe, C., Johnson, A., Jones, K.W., Jordan, N.R., Kadoya, T., Kerhoas, D., Koricheva, J., Lee, T.M., Lengyel, S., Livingstone, S.W., Lyons, A., McCabe, G., Millett, J., Montes Strevens, C., Moolna, A., Mossman, H.L., Mukherjee, N., Muñoz-Sáez, A., Negrões, N., Norfolk, O., Osawa, T., Papworth, S., Park, K.J., Pellet, J., Phillott, A.D., Plotnik, J.M., Priatna, D., Ramos, A.G., Randall, N., Richards, R.M., Ritchie, E.G., Roberts, D.L., Rocha, R., Rodríguez, J.P., Sanderson, R., Sasaki, T., Savilaakso, S., Sayer, C., Sekercioglu, C., Senzaki, M., Smith, G., Smith, R.J., Soga, M., Soulsbury, C.D., Steer, M.D., Stewart, G., Strange, E.F., Suggitt, A.J., Thompson, R.R.J., Thompson, S., Thornhill, I., Trevelyan, R.J., Usieta, H.O., Venter, O., Webber, A.D., White, R.L., Whittingham, M.J., Wilby, A., Yarnell, R.W., Zamora V., Sutherland, W.J. 2021. Training future generations to deliver evidence-based conservation and ecosystem management. Ecological Solutions and Evidence 2, e12032. https://doi.org/10.1002/2688-8319.12032

Spake, R., Mori, A.S., Beckmann, M.A., Martin, P.A., **Christie, A.P.**, Duguid, M.C., Doncaster, C.P. 2021. Implications of scale dependence for cross-study syntheses of biodiversity differences. Ecology Letters 24, 374-390. https://doi.org/10.1111/ele.13641

Junker, J., Petrovan, S.O., Arroyo-RodrÍguez, V., Boonratana, R., Byler, D., Chapman, C.A., Chetry, D., Cheyne, S.M., Cornejo, F.M., CortÉs-Ortiz, L., Cowlishaw, G., **Christie, A.P.**, Crockford, C., Torre, S.D. La, De Melo, F.R., Fan, P., Grueter, C.C., GuzmÁn-Caro, D.C., Heymann, E.W., Herbinger, I., Hoang, M.D., Horwich, R.H., Humle, T., Ikemeh, R.A., Imong, I.S., Jerusalinsky, L., Johnson, S.E., Kappeler, P.M., Kierulff, M.C.M., KonÉ, I., Kormos, R., Le, K.Q., Li, B., Marshall, A.J., Meijaard, E., Mittermeier, R.A., Muroyama, Y., Neugebauer, E., Orth, L., Palacios, E., Papworth, S.K., Plumptre, A.J., Rawson, B.M., Refisch, J., Ratsimbazafy, J., Roos, C., Setchell, J.M., Smith, R.K., Sop, T., Schwitzer, C., Slater, K., Strum, S.C., Sutherland, W.J., Talebi, M., Wallis, J., Wich, S., Williamson, E.A., Wittig, R.M., KÜhl, H.S., 2020. A Severe Lack of Evidence Limits Effective Conservation of the World's Primates. Bioscience 70, 794–803. https://doi.org/10.1093/biosci/biaa082

Campaign for Nature. 2020. A Key Sector Forgotten in the Stimulus Debate: the Nature-Based Economy. https://bit.ly/3it7F7U

Geldmann, J., Alves-Pinto, H., Amano, T., Bartlett, H., **Christie, A.P.**, Collas, L., Cooke, S.C., Correa, R., Cripps, I., Doherty, A., Finch, T., Garnett, E.E., Hua, F., Jones, J.P.G., Kasoar, T., MacFarlane, D., Martin, P.A., Mukherjee, N., Mumby, H.S., Payne, C., Petrovan, S.O., Rocha, R., Russell, K., Simmons, B.I., Wauchope, H.S., Worthington, T.A., Trevelyan, R., Green, R., Balmford, A., 2020. Insights from two decades of the Student Conference on Conservation Science. Biological Conservation 243, 108478. https://doi.org/10.1016/j.biocon.2020.108478

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., **Christie, A.P.**, Dicks, L. V, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Simmons, B.I., Balmford, A., Bladon, A.J., **Christie, A.P.**, De Palma, A., Dicks, L. V., Gallego-Zamorano, J., Johnston, A., Martin, P.A., Purvis, A., Rocha, R., Wauchope, H.S., Wordley, C.F.R., Worthington, T.A., Finch, T., 2019. Worldwide insect declines: An important message, but interpret with caution. Ecology and Evolution 9, 3678–3680. https://doi.org/10.1002/ece3.5153

Acknowledgements

I wish to thank several people, in no particular order, for their help and support during my PhD.

First, I am incredibly grateful to my supervisor, Bill Sutherland, and my two co-supervisors, Tatsuya Amano and Phil Martin, for all their wisdom, advice, and support. Bill, thank you for your dogged enthusiasm and passion for my PhD and always encouraging me to see the wider 'mad plan'. Thank you for constantly promoting my work, mentioning my name in high places, and providing me with so many opportunities to further my academic career. Tatsuya, thank you for all your insights, patience, and unwavering support that you have given me throughout my project, even when you had to relocate to the other side of the world. Phil, thank you for your advice and ideas for analyses, as well as your help in pushing all my Data Chapters through to publication.

Second, I thank my advisors, David Aldridge and Andrea Manica, for providing useful guidance to help ensure my PhD progressed smoothly. I also thank my fellow PhD students Ellie, Benno, Hannah, and Tom for their help and encouragement with all aspects of academic life. I acknowledge the Natural Environment Research Council who funded my PhD studentship through the Cambridge Earth System Science NERC DTP [NE/L002507/1].

Third, I thank all the collaborators involved with the research presented in my thesis, including Qingyuan Zhao and the numerous contributors of datasets to Chapters 2 and 3, without which many of my analyses would not have been possible. A great deal of thanks must also go to all the past and present members of the Conservation Evidence team, in particular, Rebecca Smith, Silviu Petrovan, David Williams, Claire Wordley, Katie Sainsbury, Andrew Bladon, Ricardo Rocha, Laura Pettit, Nick Littlewood, and Gorm Shackelford who contributed to the Conservation Evidence database that I analysed in Chapters 3-5. Your tireless work in summarising studies, enthusiasm for my project, and making me feel completely at home in such a friendly research group are all things I am very grateful for. I also thank the wider Conservation Science Group, as well as everyone in the Cambridge Conservation Initiative and David Attenborough Building, for their friendliness and support in such an enjoyable environment for conservation research.

Finally, I thank my fiancée, Grania, and my family for their unwavering support during my PhD. Grania, your help in reading manuscripts, listening to my boring PhD stories, advising me when problems cropped up, and helping me to cope during stressful periods of time with your humour and compassion was invaluable. To my parents, Vanessa and Mark, and brothers Scott, Euan, and Aaron, although your levels of interest in my research were positively correlated with age, your consistently high levels of support were greatly appreciated.

1 | Introduction

The importance of using evidence to inform practice and policy cannot be understated. Evidence is defined as: 'relevant information used to assess one or more hypotheses related to a question of interest' (Salafsky et al., 2019). This principle is the cornerstone of evidence-based practice and policy that seeks to provide decision-makers with reliable and relevant evidence to make better decisions (Greenhalgh, 2019; Sutherland et al., 2004). Learning from past failures and successes is a core part of using evidence as it can ultimately help to ensure future practice and policy is more effective at achieving its goals. Multiple fields, such as public healthcare, agricultural science, social science, and biodiversity conservation, are moving towards delivering the scientific evidence and syntheses that are needed to enable effective evidence-based decision-making (Alonso-Coello et al., 2016; Boutron et al., 2020; Christian et al., 2019; Donnelly et al., 2018; Kneale et al., 2019; Nakagawa et al., 2020, 2018; Porciello et al., 2020; Shackelford et al., 2019).

The origins of evidence-based practice and policy are widely attributed to the medical sciences and specifically to Archie Cochrane who drew attention to the fact that: "commonly used procedures and therapies were not always the most efficacious…" and that a substantial amount of medical practice "had not been well evaluated…" (Cochrane, 1972). His work led to what is often called an 'evidence-based revolution' in the medical sciences whereby the systematic collation of scientific evidence became more widespread and medical research became a core part of designing, adapting, and implementing medical practice and policy (Greenhalgh, 2019). The medical sciences have also served as a stark case study of why using evidence to inform practice and policy is so important. For example, many medical treatments, such as using clot-busting drugs to treat heart attacks, were only widely recommended for use decades after substantial bodies of research had been published showing this treatment was clearly beneficial (Antman et al., 1992). This underlines the importance of collating, analysing, summarising (usually termed 'evidence synthesis'), and disseminating findings from scientific research so that evidence-based practice and policy can be realised. Now many scientific fields are attempting to emulate the successes of evidence-based medicine through using evidence synthesis techniques such as systematic reviews (the systematic collation, appraisal, and summarisation of scientific evidence to answer a specific question; Collaboration for Environmental Evidence, 2013) and meta-analyses (the quantitative analysis of study findings to assess the overall estimated effect of an intervention or action across a body of scientific evidence; Gurevitch et al., 2018).

One such field is conservation science, which is facing a biodiversity crisis as human-driven threats are extirpating species and habitats at an alarming rate (IPBES secretariat, 2019;

Leclère et al., 2020). The need for evidence-based conservation in biodiversity conservation is particularly acute given the lack of time left to save many species from extinction and the underfunded and resource-limited nature of conservation efforts (Sutherland et al., 2004; Williams et al., 2020). One major rationale behind evidence-based decision-making is to make practice and policy more efficient and effective, preventing the waste of time and resources on implementing interventions that are known to be inefficient or ineffective. Evidence-based conservation is still relatively early in its progress towards making evidence-based decision-making common practice in conservation, but important strides have been made towards this goal in the past few decades.

One of the major developments in realising evidence-based conservation is the more widespread collation of scientific evidence using standardised methods of evidence synthesis (Nakagawa et al., 2020; Salafsky et al., 2019). The most widely used of these are systematic reviews and meta-analyses. Systematic reviews, within which quantitative meta-analyses may be conducted, typically have a narrow scope and focus on a clearly defined question or set of questions (e.g., how does tillage intensity affect soil organic carbon? Haddaway et al., 2017). Publishing and peer-reviewing the protocols underlying systematic reviews is widely recommended and an important precursor to conducting such a review. There is still some progress to be made to ensure systematic reviews are indeed systematic, reduce their susceptibility to biases, and use rigorous approaches to collating and assessing scientific evidence (Haddaway et al., 2020). Systematic maps are another widely used tool for evidence synthesis (Bates et al., 2007). These maps of evidence do not aim to answer questions, unlike systematic reviews, but instead aim to describe and summarise the state of the evidence that is available to answer a question, particularly in terms of the quantity, quality, and distribution of evidence (e.g., Fagerholm et al., 2016). Systematic maps often act as a precursor to systematic reviews, assessing whether there is sufficient evidence available to meaningfully answer a certain question and providing recommendations for future studies that could help fill gaps in the evidence base (Collaboration for Environmental Evidence, 2013).

Another important method of evidence synthesis addresses the challenge of collating, summarising, and assessing evidence from a different angle. This is called subject-wide evidence synthesis (Sutherland et al., 2019) and is primarily used by the Conservation Evidence project (Conservation Evidence, 2020). A subject, defined here, is an area such as bird conservation, whilst a discipline is considered a broader topic such as biodiversity conservation. Subject-wide evidence synthesis involves conducting discipline-wide literature searches in core conservation literature sources (e.g., academic journals, report series, organisational websites), using defined protocols and study inclusion criteria, to collate relevant studies that quantitatively test conservation interventions into a large database.

These core literature sources are identified based on their likelihood to contain relevant information on conservation interventions. Subject-specific literature sources are then searched, depending on the subject for which evidence is being synthesised, to supplement searches of core conservation journals (Sutherland et al., 2019). Studies are assigned to the conservation interventions they test, and the context and findings of studies are summarised in plain language. The studies for each intervention are assessed as a whole and summarised using key messages, whilst an expert panel uses a modified Delphi technique (Sutherland et al., 2019) to score the effectiveness, certainty, and harms of each intervention (see www.conservationevidence.com). The scope for this form of evidence synthesis is deliberately broader than systematic reviews because the aim is to collate, summarise and assess the evidence base for entire subjects (e.g., amphibian or bird conservation), answering multiple questions about the effectiveness of different conservation interventions simultaneously (Sutherland et al., 2019). This approach is, however, extremely costly in the short-term because searches are not narrow in scope and require the searching of literature for entire disciplines on broad subjects. The aim is, however, to maximise the long-term efficiency of evidence synthesis through benefitting from economies of scale (Sutherland et al., 2019); once studies are included in the database, the database can be dynamically updated with newly published studies and the evidence for different interventions can be assessed all at once without needing to conduct repeat searches (and effectively repeat a time-consuming systematic review; Haddaway and Westgate, 2019). In certain ways, subject-wide evidence synthesis also performs the role of a systematic map, highlighting gaps in the evidence base and showing the distribution, quantity, and quality of evidence available for informing practice and policy (Sutherland et al., 2019). This methodology also allows decision-makers to view digestible summaries of evidence that are open access and freely available – and so helps to remove barriers to the access of evidence (e.g., paywalls; Fuller et al. 2014; Sunderland et al. 2009).

Whilst these different approaches to evidence synthesis have led to improvements in the accessibility, availability, and digestibility of the evidence base for conservation, evidence-based conservation still faces several challenges. In this thesis, I aim to quantify and address fundamental challenges to achieving more widespread and successful evidence-based decision-making in biodiversity conservation practice and policy. I focus on the challenges that relate specifically to the biases and knowledge gaps present in the evidence base for conservation. In line with its title, this thesis is structured into quantifying and addressing two types of biases that exist in any scientific evidence base: within-study and between-study biases (Fig.1).

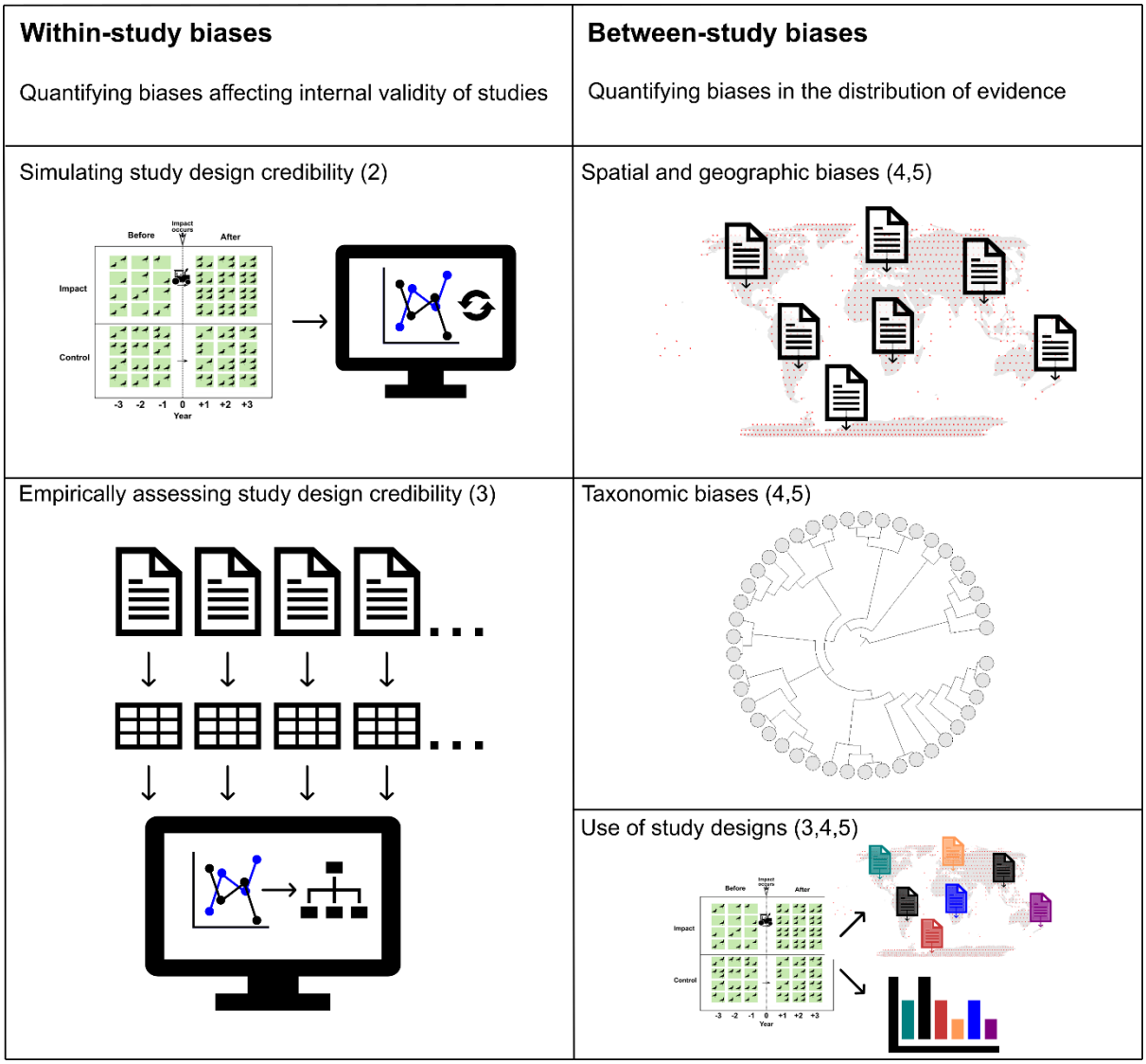


Figure 1 *– A diagrammatic outline of the structure of this thesis to quantify and address two types of biases in the evidence base for conservation: within-study and between-study biases. Chapter numbers are given in parentheses alongside a brief description of the approaches used or biases that are tackled.*

I define within-study biases as those affecting the reliability of research findings (i.e., internal validity or the extent to which the evidence can be reasonably relied upon to form an opinion or inference; Weed, 2005), which in turn is known to hinge upon the choice of study design used to collect data. Many study designs are used to test the effectiveness of interventions, from gold standard randomised experiments (in the medical sciences) to various types of observational designs that are often used when randomised experiments are too expensive, unfeasible, or unethical to implement (Angrist and Pischke, 2008; Imbens and Rubin, 2015; Rubin, 2008). Whilst there is generally a good qualitative understanding of the relative advantages and disadvantages of different study designs and the biases affecting them, we lack a thorough quantitative understanding of how much trust should be placed in findings

obtained using one design over another. This is a fundamental question that is relevant to evidence-based decision-making as study quality and uncertainty is a major factor that must be incorporated when assessing the strength of scientific evidence (Bilotta et al., 2014; Mupepele et al., 2016; Shea et al., 2017). Whilst between-study and small within-study comparisons of the results obtained using two different designs are common in medicine and the social sciences (e.g., Altindag et al., 2019; Chaplin et al., 2018; Cook et al., 2008), large-scale within-study comparisons of different study designs are rare and have not been conducted in the environmental sciences.

I tackle the issue of within-study design biases by quantitatively estimating the relative reliability of results obtained by different commonly used study designs in the environmental sciences. In Chapter 2, I simulate the performance of different study designs using empirically derived estimates of the magnitude of study design bias from 51 ecological datasets. These datasets were obtained from a range of studies around the world in different ecological fields. In Chapter 3, I build on these simulations by digging deeper into the raw datasets, conducting pairwise comparisons of the estimates given by different designs within each dataset. I also develop a hierarchical Bayesian model to quantify the relative reliability of study designs, enabling meta-analyses to better account for the bias and variance introduced by studies. This approach attempts to tackle the challenging issue of combining study results obtained using different study designs, which has been a hotly debated issue in evidence synthesis. Insights from Chapters 2 and 3 therefore provide recommendations on how to improve the quality and reliability of the evidence base for conservation, as well as ways to improve how evidence synthesis accounts for uncertainty and biases within studies.

Understanding between-study biases, or biases affecting the wider literature's distribution and coverage, is crucial to prioritising future conservation research and action and understanding the relevance of evidence to a decision-maker. Relevance can be defined as the extent to which any single piece of evidence could have the tendency to make a fact more or less probable (Weed, 2005). Scientific evidence is typically synthesised at a global level and therefore products of evidence synthesis (e.g., systematic maps and reviews) have been criticised in the past as offering a 'view from nowhere' (Shapin, 1998), lacking a consideration of relevance or realism (see Levins, 1966, 1968) that helps to provide decision-makers with locally valid and useful evidence-based recommendations (Gutzat and Dormann, 2020). In conservation, practitioners have been shown to prefer locally valid evidence (high relevance to their local context) since the complexity of ecosystems is perceived to make generalising difficult (Gutzat and Dormann, 2020). This evidence may often come in the form of 'local knowledge' (e.g., Local Ecological Knowledge (LEK) or tacit knowledge), based on the experience or intuition of practitioners, stakeholders, and decision-makers (Tanner et al.,

2020; Wheeler and Root-Bernstein, 2020), which may not be reflected in evidence syntheses of the scientific literature. Whilst this local knowledge is useful and represents highly relevant evidence, its reliability may be poor and difficult to verify, and may not be disseminated to improve practice (Dicks et al. 2014).

A key motivation for the movement towards Evidence-Based Conservation arose from the continued use of actions by practitioners which had been previously shown to be ineffective elsewhere, often in the scientific literature, as well as the perceived overreliance of practitioners on their own knowledge to inform their management actions (often without consulting scientific evidence from different, but ecologically relevant habitats and species, for example; Sutherland et al. 2004). Nevertheless, a consensus is now growing that combining evidence derived from both the scientific literature and local knowledge is required to reach meaningful and effective evidence-informed decisions (e.g., see Cook et al. 2013; Kadykalo et al. 2021a,b). To combine these forms of evidence, it is important to understand the relevance (or 'external validity' or applicability) of scientific findings and whether the needs of practitioners for locally valid evidence are fulfilled by the current state of the scientific evidence base for conservation.

I start to assess these issues in Chapter 4, using the database from the Conservation Evidence project containing thousands of quantitative tests of conservation interventions, I quantify the spatial, taxonomic, bioclimatic, and design-related biases to show the severity of the scientific knowledge gaps for amphibian and bird conservation. I also briefly quantify the use of study designs in the environmental and social sciences in Chapter 3 as they provide important context to this Chapter's findings. In Chapter 5, I consider the utility of the evidence base for conservation for informing local action. As discussed earlier, this is important because decision-makers that we, as scientists, try to inform typically prefer evidence that is locally valid and relevant to their specific context (Gutzat and Dormann, 2020). I quantify how much locally relevant evidence exists for decision-makers and whether priorities for testing interventions have previously been aligned with the urgency and need for conservation interventions for different species and locations. Insights from Chapters 4 and 5 will help to improve and understand the relevance of scientific evidence, as well as to prioritise future research effort to address between-study biases in the evidence base for conservation.

Chapter 6 discusses the future of evidence-based conservation and provides recommendations, based on the findings of this thesis, on how to ensure meaningful evidence-based decision-making is ultimately realised in biodiversity conservation.

References

Alonso-Coello, P., Oxman, A.D., Moberg, J., Brignardello-Petersen, R., Akl, E.A., Davoli, M., Treweek, S., Mustafa, R.A., Vandvik, P.O., Meerpohl, J., Guyatt, G.H., Schünemann, H.J., 2016. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. British Medical Journal 353, i2089. https://doi.org/10.1136/bmj.i2089

Altindag, O., Joyce, T.J., Reeder, J.A., 2019. Can Nonexperimental Methods Provide Unbiased Estimates of a Breastfeeding Intervention? A Within-Study Comparison of Peer Counseling in Oregon. Evaluation Review 43, 152–188. https://doi.org/10.1177/0193841X19865963

Angrist, J.D., Pischke, J.-S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F., Chalmers, T.C., 1992. A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction. JAMA 268, 240–248. https://doi.org/10.1001/jama.1992.03490020088036

Bates, S., Clapton, J., Coren, E., 2007. Systematic maps to support the evidence base in social care. Evidence and Policy: A Journal of Research, Debate and Practice 3, 539–551.

Bilotta, G.S., Milner, A.M., Boyd, I.L., 2014. Quality assessment tools for evidence from environmental science. Environmental Evidence 3, 14. https://doi.org/10.1186/2047-2382-3-14

Boutron, I., Créquit, P., Williams, H., Meerpohl, J., Craig, J.C., Ravaud, P., 2020. Future of evidence ecosystem series: 1. Introduction — Evidence synthesis ecosystem needs dramatic change. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2020.01.024

Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N., Morris, R.E., 2018. The Internal And External Validity Of The Regression Discontinuity Design: A Meta-Analysis Of 15 Within-Study Comparisons. Journal of Policy Analysis and Management 37, 403–429. https://doi.org/10.1002/pam.22051

Christian, D., Amano, T., González-varo, J.P., Mukherjee, N., Robertson, R.J., Simmons, B.I., Wauchope, H.S., Sutherland, W.J., 2019. Calling for a new agenda for conservation science to create evidence- informed policy. Biological Conservation 238, 108222. https://doi.org/10.1016/j.biocon.2019.108222

Cochrane, A., 1972. Effectiveness and Efficiency: Random Reflections on Health Services, Nuffield Trust.

Collaboration for Environmental Evidence, 2013. Guidelines for Systematic Reviews in Environmental Management. Environmental Evidence 4.2, 80. https://doi.org/www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf

Conservation Evidence, 2020. Conservation Evidence [WWW Document]. URL www.conservationevidence.com (accessed 2.1.21).

Cook, T.D., Shadish, W.R., Wong, V.C., 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management 27, 724–750. https://doi.org/10.1002/pam.20375

Cook, C.N., Mascia, M.B., Schwartz, M.W., Possingham, H.P., Fuller, R.A., 2013. Achieving Conservation Science that Bridges the Knowledge–Action Boundary. Conservation Biology 27, 669–678. https://doi.org/10.1111/cobi.12050

Dicks, L.V., Walsh, J.C., Sutherland, W.J., 2014. Organising evidence for environmental management decisions: A '4S' hierarchy. Trends in Ecology & Evolution 29, 607–613. https://doi.org/10.1016/j.tree.2014.09.004.

Donnelly, C.A., Boyd, I., Campbell, P., Craig, C., Vallance, P., Walport, M., Whitty, C.J.M., Woods, E., Wormald, C., 2018. Four principles to make evidence synthesis more useful for policy. Nature 558, 361–364. https://doi.org/10.1038/d41586-018-05414-4

Fagerholm, N., Torralba, M., Burgess, P.J., Plieninger, T., 2016. A systematic map of ecosystem services assessments around European agroforestry. Ecological Indicators 62, 47–65. https://doi.org/https://doi.org/10.1016/j.ecolind.2015.11.016

Fuller, R.A., Lee, J.R., Watson, J.E.M., 2014. Achieving open access to conservation science. Conservation Biology 28, 1550–1557. https://doi.org/10.1111/cobi.12346

Greenhalgh, T., 2019. How to read a paper: the basics of Evidence Based Medicine, 6th ed. John Wiley & Sons Ltd., Hoboken.

Gurevitch, J., Koricheva, J., Nakagawa, S., Stewart, G., 2018. Meta-analysis and the science of research synthesis. Nature 555, 175–182. https://doi.org/10.1038/nature25753

Gutzat, F., Dormann, C.F., 2020. Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice. Environmental Management 66, 435–449. https://doi.org/10.1007/s00267-020-01312-6

Haddaway, N.R., Bethel, A., Dicks, L. v, Koricheva, J., Macura, B., Petrokofsky, G., Pullin, A.S., Savilaakso, S., Stewart, G.B., 2020. Eight problems with literature reviews and how to fix them. Nature Ecology and Evolution. https://doi.org/10.1038/s41559-020-01295-x

Haddaway, N.R., Hedlund, K., Jackson, L.E., Kätterer, T., Lugato, E., Thomsen, I.K., Jørgensen, H.B., Isberg, P.-E., 2017. How does tillage intensity affect soil organic carbon? A systematic review. Environmental Evidence 6, 30. https://doi.org/10.1186/s13750-017-0108-9

Haddaway, N.R., Westgate, M.J., 2019. Predicting the time needed for environmental systematic reviews and systematic maps. Conservation Biology 33, 434–443. https://doi.org/https://doi.org/10.1111/cobi.13231

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge.

IPBES secretariat, 2019. Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

Kadykalo, A.N., Buxton, R.T., Morrison, P., Anderson, C.M., Bickerton, H., Francis, C.M., Smith, A.C., Fahrig, L., 2021a. Bridging research and practice in conservation. Conservation Biology. Accepted Author Manuscript. https://doi.org/10.1111/cobi.13732

Kadykalo, A.N., Cooke, S.J., Young, N., 2021b. The role of western-based scientific, Indigenous and local knowledge in wildlife management and conservation. People and Nature. Accepted Author Manuscript. https://doi.org/10.1002/pan3.10194

Kneale, D., Thomas, J., O'Mara-Eves, A., Wiggins, R., 2019. How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision making? Research Synthesis Methods 10, 44–56. https://doi.org/10.1002/jrsm.1320

Leclère, D., Obersteiner, M., Barrett, M., Butchart, S.H.M., Chaudhary, A., de Palma, A., DeClerck, F.A.J., di Marco, M., Doelman, J.C., Dürauer, M., Freeman, R., Harfoot, M., Hasegawa, T., Hellweg, S., Hilbers, J.P., Hill, S.L.L., Humpenöder, F., Jennings, N., Krisztin, T., Mace, G.M., Ohashi, H., Popp, A., Purvis, A., Schipper, A.M., Tabeau, A., Valin, H., van Meijl, H., van Zeist, W.-J., Visconti, P., Alkemade, R., Almond, R., Bunting, G., Burgess, N.D., Cornell, S.E., di Fulvio, F., Ferrier, S., Fritz, S., Fujimori, S., Grooten, M., Harwood, T., Havlík, P., Herrero, M., Hoskins, A.J., Jung, M., Kram, T., Lotze-Campen, H., Matsui, T., Meyer, C., Nel, D., Newbold, T., Schmidt-Traub, G., Stehfest, E., Strassburg, B.B.N., van Vuuren, D.P.,

Ware, C., Watson, J.E.M., Wu, W., Young, L., 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. Nature 585, 551–556. https://doi.org/10.1038/s41586-020-2705-y

Levins, R., 1968. Evolution in changing environments: some theoretical explorations. Princeton University Press, New Jersey.

Levins, R., 1966. The strategy of model building in population biology. American scientist 54, 421–431. https://www.jstor.org/stable/27836590

Mupepele, A.-C., Walsh, J.C., Sutherland, W.J., Dormann, C.F., 2016. An evidence assessment tool for ecosystem services and conservation studies. Ecological Applications 26, 1295–1301. https://doi.org/10.1890/15-0595

Nakagawa, S., Dunn, A.G., Lagisz, M., Bannach-Brown, A., Grames, E.M., Sánchez-Tójar, A., O'Dea, R.E., Noble, D.W.A., Westgate, M.J., Arnold, P.A., Barrow, S., Bethel, A., Cooper, E., Foo, Y.Z., Geange, S.R., Hennessy, E., Mapanga, W., Mengersen, K., Munera, C., Page, M.J., Welch, V., Haddaway, N.R., 2020. A new ecosystem for evidence synthesis. Nature Ecology and Evolution 4, 498–501. https://doi.org/10.1038/s41559-020-1153-2

Nakagawa, S., Samarasinghe, G., Haddaway, N.R., Westgate, M.J., O'Dea, R.E., Noble, D.W.A., Lagisz, M., 2019. Research Weaving: Visualizing the Future of Research Synthesis. Trends in Ecology and Evolution 34, 224–238. https://doi.org/10.1016/j.tree.2018.11.007

Porciello, J., Ivanina, M., Islam, M., Einarson, S., Hirsh, H., 2020. Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. Nature Machine Intelligence 2, 559–565. https://doi.org/10.1038/s42256-020-00235-5

Rubin, D.B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2, 808–840. https://doi.org/10.1214/08-AOAS187

Salafsky, N., Boshoven, J., Burivalova, Z., Dubois, N.S., Gomez, A., Johnson, A., Lee, A., Margoluis, R., Morrison, J., Muir, M., Pratt, S.C., Pullin, A.S., Salzer, D., Stewart, A., Sutherland, W.J., Wordley, C.F.R., 2019. Defining and using evidence in conservation practice. Conservation Science and Practice 1, e27. https://doi.org/10.1111/csp2.27

Shackelford, G.E., Kelsey, R., Sutherland, W.J., Kennedy, C.M., Wood, S.A., Gennet, S., Karp, D.S., Kremen, C., Seavy, N.E., Jedlicka, J.A., Gravuer, K., Kross, S.M., Bossio, D.A., Muñoz-Sáez, A., LaHue, D.G., Garbach, K., Ford, L.D., Felice, M., Reynolds, M.D., Rao, D.R., Boomer, K., LeBuhn, G., Dicks, L. v, 2019. Evidence Synthesis as the Basis for Decision Analysis: A Method of Selecting the Best Agricultural Practices for Multiple Ecosystem

Services. Frontiers in Sustainable Food Systems 3, 83. https://doi.org/10.3389/fsufs.2019.00083

Shapin, S., 1998. Placing the view from nowhere: historical and sociological problems in the location of science. Transactions of the Institute of British Geographers 23, 5–12. https://doi.org/10.1111/j.0020-2754.1998.00005.x

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., Henry, D.A., 2017. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ (Online) 358, 1–8. https://doi.org/10.1136/bmj.j4008

Sunderland, T., Sunderland-Groves, J., Shanley, P., Campbell., B., 2009. Bridging the gap: How can information access and exchange between conservation biologists and field practitioners be improved for better conservation outcomes? Biotropica 41, 549–554. https://doi.org/10.1111/j.1744-7429.2009.00557.x

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19, 305–308. https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Tanner, L., Mahajan, S.L., Becker, H., DeMello, N., Komuhangi, C., Mills, M., Masuda, Y., Wilkie, D., Glew, L., 2020. Making better decisions: How to use evidence in a complex world. The Research People and the Alliance for Conservation Evidence and Sustainability. https://www.allianceconservationevidence.org/s/Making_better_decisions_ACES.pdf

Weed, D.L., 2005. Weight of Evidence: A Review of Concept and Methods. Risk Analysis 25, 1545–1557. https://doi.org/10.1111/j.1539-6924.2005.00699.x

Wheeler, H.C., Root-Bernstein, M., 2020. Informing decision-making with Indigenous and local knowledge and science. Journal of Applied Ecology 57, 1634–1643. https://doi.org/10.1111/1365-2664.13734

Williams, D.R., Balmford, A., Wilcove, D.S., 2020. The past and future role of conservation science in saving biodiversity. Conservation Letters 13, 1–7. https://doi.org/10.1111/conl.12720

2 | Simple study designs in ecology produce inaccurate estimates of biodiversity responses

A modified version of this chapter was published as:

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Abstract

1. Monitoring the impacts of anthropogenic threats and interventions to mitigate these threats is key to understanding how to best conserve biodiversity. Ecologists use many different study designs to monitor such impacts. Simpler designs lacking controls (e.g., Before-After (BA) and After) or pre-impact data (e.g., Control-Impact (CI)) are considered to be less robust than more complex designs (e.g., Before-After Control-Impact (BACI) or Randomised Control Impact (RCI)). However, we lack quantitative estimates of how much less accurate simpler study designs are in ecology. Understanding this could help prioritise research and weight studies by their design's accuracy in meta-analysis and evidence assessment.
2. We compared how accurately five study designs estimated the true effect of a simulated environmental impact that caused a step-change response in a population's density. We derived empirical estimates of several simulation parameters from 47 ecological datasets to ensure our simulations were realistic. We measured design performance by determining the percentage of simulations where: (i) the true effect fell within the 95% Confidence Intervals of effect size estimates, and (ii) each design correctly estimated the true effect's direction and magnitude. We also considered how sample size affected their performance.
3. We demonstrated that RCI and BACI designs are far more accurate than BA, CI and After designs. When estimating the true effect to within ±30% and correctly identifying its direction (in terms of statistical significance), RCI performed (depending on sample size) 2.5-2.6 times better than BA, 2.6-2.8 times than CI, and 6.1-7.3 times than After designs, whilst BACI performed 1.6-2.0 times better than BA, 1.8-2.2 times than CI, and 4.8-5.2 times than After designs. By this measure, RCI also performed approximately 1.1-1.3 times better than BACI. RCI and BACI designs suffered from low statistical power at small sample sizes, but still outperformed the other simpler designs for a range of performance measures. Increasing sample size only increased precision in simpler designs (CI, BA, and After) around a more biased estimate of the true effect compared to RCI and BACI designs.
4. *Synthesis and applications.* We suggest that more investment in more robust designs is needed in ecology since inferences from simpler designs, even with large sample sizes may be misleading. Facilitating this requires longer-term funding and stronger research-practice partnerships. We also propose 'accuracy weights' and demonstrate how they can weight studies in three recent meta-analyses by accounting for study design and sample size. We hope these help decision-makers and meta-analysts better account for study design when assessing evidence.

Introduction

Monitoring the impact of human activities on biodiversity is fundamental to understanding how to effectively conserve biodiversity. This includes monitoring the impacts of anthropogenic threats, as well as the effectiveness of management interventions to mitigate such threats. The main challenge for such monitoring is disentangling natural environmental change from anthropogenic change (Hewitt et al., 2001; Hipel et al., 1978), whilst considering the focal impact's statistical (Box and Tiao, 1975; Osenberg and Schmitt, 1996) and ecological significance (Wolfe et al., 1987). The complexity of ecosystems, including various sources of spatiotemporal variation and confounding variables, has catalysed much research on understanding the best ways to design impact assessments (Hipel et al., 1978; Lettenmaier et al., 1978; Osenberg et al., 2006; Stewart-Oaten et al., 1986). Whilst improvements in study design have helped ecologists to quantify human impacts on biodiversity more accurately, a range of designs with varying complexity and biases still persist (de Palma et al., 2018; Table 1).

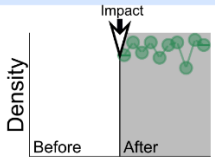
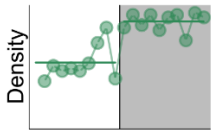
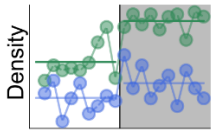
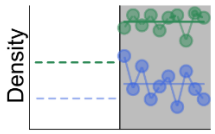
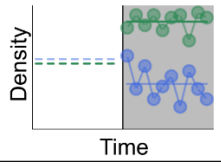
Study design is composed of three major aspects: (i) pre-impact sampling, (ii) use of controls, and (iii) randomised allocation of independent sampling units (here we term these "sites"). Adding pre-impact sampling to an After design – where monitoring only occurs after the impact – produces the Before-After (BA) design (Table 1). This compares the system's state before and after the impact, attempting to minimise bias from temporal variability and pre-impact conditions.

Addition of control sites to BA designs results in Before-After Control-Impact (BACI) designs, where the average difference between control and impact sites is compared before and after an impact (Table 1; Osenberg et al., 2006; Stewart-Oaten et al., 1986). BACI designs use the pre-impact differences between control and impact sites as a null hypothesis for post-impact differences that would exist if the focal impact were absent – avoiding bias from a lack of a control (Thiault et al., 2017). Problems with site-specific temporal variation in BACI designs can be addressed by sampling control and impact sites simultaneously, several times before and after the impact (Before-After Control-Impact Paired-Series (BACIPS) design; Stewart-Oaten and Bence, 2001).

Random allocation of sites to control and impact groups represents the third major aspect of study design. Control-Impact (CI) designs, analogous with Space-For-Time Substitutions (de Palma et al., 2018; França et al., 2016) or Intervention Versus Reference Site designs (Stewart-Oaten and Bence, 2001), compare non-randomly allocated control and impact sites after the impact (Table 1). However, this non-random allocation can violate the assumption that the only differences between control and impact sites are due to the focal impact, leading

to biased results (Damgaard, 2019; de Palma et al., 2018; Larsen et al., 2019; Table 1). Randomised Control-Impact (RCI) minimise this bias by randomising site allocation to impact and control groups (Table 1). This reduces the need to sample before and after the impact to account for any initial differences (i.e., as in the BACI design) if sufficient numbers of sites and points in time are sampled (de Palma et al., 2018; Larsen et al., 2019).

Table 1 *– Comparison of the key features of study designs. Graphs show how designs sample from impact (green points) and control (blue points) sites over time, before and after an impact (white versus grey areas, respectively). Solid horizontal lines show the average density of sites measured to calculate each design's effect size estimate. Dashed horizontal lines for CI and RCI represent the pre-impact differences between the mean densities of control and impact sites, which cause bias for CI and noise for RCI. Many design variants exist – e.g., MBACI for BACI with multiple sites, R for Reference in BARI (Webb et al., 2012).*

Design	Sampling regime	Relative cost	Relative difficulty in ecology	Applicability	Ecological examples of use
After		Very low	Very low	• Most systems • Where control unfeasible • Unpredictable impacts	Pond creation (e.g., Merrow, 2007).
Before-After (BA)		Moderate	Moderate	• Predictable impacts • Where control unfeasible • Availability of pre-impact data	Wildlife tunnels under roads (e.g., Scoccianti, 2006).
Before-After Control-Impact (BACI) (BARI, MBACI, BACIPS)		High	High	• Predictable impacts • Appropriate control • Availability of pre-impact data • Parallel trends prior to impact	MPA effectiveness, mitigation of renewable energy infrastructure (e.g., Fernández-Chacón et al., 2015).
Control-Impact (CI) (Space-for-Time, Impact versus Reference Sites)		Low	Moderate	• Unpredictable impacts • Large-scale replicates that cannot be truly randomised	Monitoring or mitigation of oil spill or other pollution event (e.g., Westgate et al., 1998).
Randomised Control-Impact (RCI) (Randomised Controlled Trial, RCT)		Low	Very high	• Unpredictable impacts • Small-scale replicates appropriate for randomisation	Peatland restoration, creation of field margins (e.g., Huusela-Veistola, 1998).

Despite the development of robust approaches to quantifying impacts, greater use of less robust designs persists. Three systematic maps on the biodiversity impacts of different threats and interventions found that a low proportion of studies used BACI (6-29%) and BA designs (3-37%), but many more used CI designs (48-89%) (Bernes et al., 2017, 2015; Papathanasopoulou et al., 2016).

The greater prevalence of CI designs in the ecological literature probably reflects that they can be easier to implement than more complex study designs. For example, RCI is widely used in fields, such as medicine, where random allocation of small-scale experimental units to impact and control groups is possible (Downs and Black, 1998; Tugwell and Haynes, 2006). However, RCI often cannot be used in ecology because true randomisation of experimental units is more difficult with large-scale sites (e.g., protected areas) compared to smaller, more readily available plots (Larsen et al., 2019; Stewart-Oaten and Bence, 2001). Therefore, ecologists tend to use pseudo-experimental designs lacking randomisation, such as BA, CI, and BACI designs (Table 1; de Palma et al., 2018). Nevertheless, constraints due to cost, logistics, and project duration often prevent the implementation of complex BACI and even simpler BA designs because of the need to revisit sites pre- and post-impact (França et al., 2016; Osenberg et al., 2011); Table 1).

The disparities between the robustness of study designs and their usage are concerning as many studies may be making misleading inferences about anthropogenic impacts. Some empirical comparisons of the consequences of using BACI, BA, and CI designs have been undertaken (França et al., 2016; Mahlum et al., 2018; Osenberg et al., 2011; Smokorowski et al., 2017). However, we are yet to understand how inaccurate simpler designs are relative to complex ones, or the influence of sample size on these patterns (e.g., are simpler designs with large sample sizes equivalent to more complex designs with smaller sample sizes?). A quantitative comparison of the accuracy of different designs and their sample size would help us better understand these issues.

To address this knowledge gap, we simulate a hypothetical population's response to an impact, and compare how accurately different study designs estimate that response. We use empirically derived parameter estimates from 47 ecological datasets to generate realistic control and impact data, before and after an impact. BACI, RCI, BA, CI and After designs are then used to sample from the simulated data with various levels of spatial replication (control and impact sites). We compare the accuracy of each design by their ability first to predict the correct direction of the response, and second to estimate the response to within a given percentage. Our goal is to inform the development of a quantitative scale of the comparative accuracy of different designs. Such a scale would have utility for future monitoring of anthropogenic impacts, as well as assessing the quality of ecological studies used to inform policy and practice.

Materials and methods

We simulated a hypothetical population with true density λ that varied over T time steps before and after a chronic impact occurred (Fig.1). For example, if T=10, then time steps 1-10 were classified as the 'before period' (i.e., before the impact occurred) and time steps 11-20 were classified as the 'after period' (Fig.1). The true density was monitored in sites where the impact occurred ('impact sites') and where the impact was absent ('control sites').

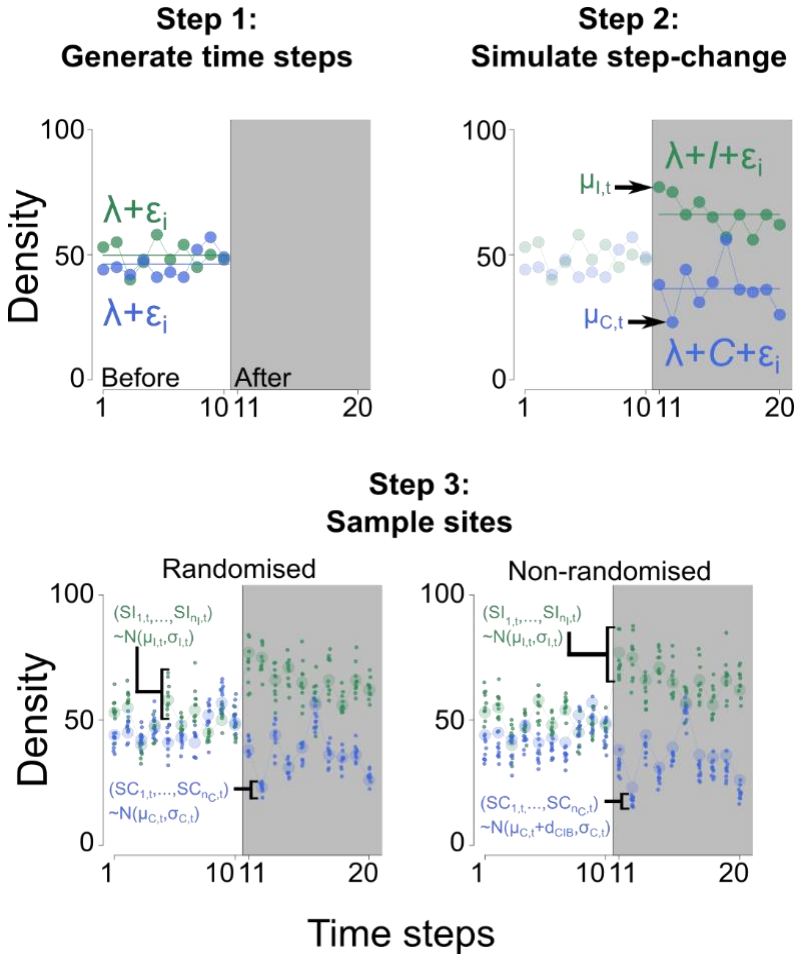


Figure 1 *– An overview of our simulation. Step 1 shows true densities of control and impact sites generated in the before period (white area). Step 2 shows true densities of control and impact sites generated in the after period (grey area) to reflect a step-change response (using I and C); the true density in each time step (t) is shown ($\mu_{I,t}$, impact: green; and $\mu_{C,t}$, control: blue). Step 3 shows how control and impact sites (SI and SC) are sampled (n_I and n_C = 10) for both randomised and non-randomised designs.* R code to replicate the simulation is available via Zenodo: https://doi.org/10.5281/zenodo.4437010.

We set the mean true density to 50 and randomly sampled T values from a Poisson distribution (λ=50) to vary the true density over T time steps in the before period for control and impact sites. These T values defined the true density in each time step before the impact occurred

(e.g., $\mu_{I,t}$for impact sites in the *t*th time step). To simulate a step-change response at both control and impact sites after the impact occurred (Fig.1), we sampled from a different Poisson distribution with λ adjusted by an empirically derived amount I (λ+ I; Fig.1; Table 2) for impact sites and an empirically derived amount C for control sites (λ+ C; Fig.1; Table 2). I and C were varied using empirical estimates of the proportional change in control and impact sites in the before period versus the after period, p_I and p_C, respectively ($I = \lambda \cdot (p_I - 1)$; $C = \lambda \cdot (p_C - 1)$; Table 2), derived from 47 ecological datasets from a range of different locations and disciplines (see Appendix S1 for details of how these were obtained, and Appendix S5 for a list of published data sources containing these datasets).

Whilst we focus on a step-change response in our simulation, temporal biodiversity dynamics following disturbances or interventions can follow different trajectories (di Fonzo et al., 2013; Thiault et al., 2017). However, to simplify the simulation as much as possible, particularly in terms of computational demands, using a step-change response was most appropriate to test the relative accuracy of each design.

Using the simulated data for before and after periods we sampled various numbers of impact (n_I) and control (n_C) sites (these could also be plots or transects; Fig.1). For RCI that uses random allocation of sites to control and impact groups, we randomly sampled sites from two normal distributions for each time step: one with a mean, $\mu_{I,t}$, for impact sites and one with a mean, $\mu_{C,t}$, for control sites (Fig.1). The number of sites sampled was the same for all time steps. The standard deviation of each normal distribution represented the variation amongst sites and was calculated by multiplying the mean by the coefficient of variation (e.g., control sites: $\sigma_{C,t} = \mu_{C,t} \cdot CV_S$; impact sites: $\sigma_{I,t} = \mu_{I,t} \cdot CV_S$; Table 2). We varied CV_S by randomly drawing values from a truncated normal distribution: N($\mu = 0.1, \sigma = 0.05, min = 0, max = 0.2$).

Table 2 *– Definitions and summary statistics for all simulation parameters (termed 'Sim.') and empirically derived parameters (termed 'Emp.'; Appendix S1). Equations show how each parameter was calculated. For empirically derived parameters, $\bar{x}$ refers to the average of sampled sites taken from 47 ecological datasets (e.g., $\overline{x_{AC}}$ refers to the average of all control sites in the after period; Appendix S1).*

Para-meter	Definition	Source	Equation	Mean	SD	Min	Max
p_C	Change in control between before and after periods	Emp.	$p_C = \frac{\lvert\overline{x_{AC}}\rvert}{\lvert\overline{x_{BC}}\rvert}$	0.918	0.181	0.605	1.31
p_I	Change in impact between before and after periods	Emp.	$p_I = \frac{\lvert\overline{x_{AI}}\rvert}{\lvert\overline{x_{BI}}\rvert}$	0.967	0.230	0.579	1.46
p_{CIB}	Average value of control sites as a proportion of the average value of impact sites in the before period	Emp.	$p_{CIB} = \frac{\lvert\overline{x_{BC}}\rvert}{\lvert\overline{x_{BI}}\rvert}$	1.13	0.306	0.654	1.89
I	True change in impact sites from before to after impact	Emp.	$I = \lambda \cdot (p_I - 1)$	-1.65	11.5	-21.1	23.2
C	True change in control sites from before period to after period	Emp.	$C = \lambda \cdot (p_C - 1)$	-4.10	9.05	-19.8	15.4
d_{CIB}	Difference between true densities of control and impact sites in before period	Emp.	$d_{CIB} = \lambda \cdot (p_{CIB} - 1)$	6.60	15.3	-17.3	44.5
λ	True density across all time steps	Sim.	$\lambda = 50$	-	-	50	50
T	Total number of time steps simulated	Sim.	$T = \{2,4,6,8,10\}$	-	-	2	10
n_T	Number of time steps sampled in each period	Sim.	$n_T = T$	-	-	2	10
$\mu_{I,t}$	True density in impact sites in time step t	Sim.	Before: $\mu_{I,t} \sim Poisson(\lambda)$ After: $\mu_{I,t} \sim Poisson(\lambda + I)$	-	-	-	-
$\sigma_{I,t}$	Standard deviation of impact sites in time step t	Sim.	$\sigma_{I,t} = CV_S \cdot \mu_{I,t}$	-	-	-	-
$\mu_{C,t}$	True density in control sites in time step t	Sim.	Before: $\mu_{C,t} \sim Poisson(\lambda)$ After: $\mu_{C,t} \sim Poisson(\lambda + C)$	-	-	-	-
$\sigma_{C,t}$	Standard deviation of control sites in time step t	Sim.	$\sigma_{C,t} = CV_S \cdot \mu_{C,t}$	-	-	-	-
CV_S	Coefficient of variation (variation amongst sites)	Sim.	$CV_S \sim N(\mu, \sigma, min, max)$	0.10	0.05	0.00	0.20
$SI_{n,t}$	nth impact site sampled in time step t	Sim.	$(SI_{1,t},\ldots,SI_{nI,t}) \sim N(\mu_{I,t}, \sigma_{I,t})$	-	-	-	-
$SC_{n,t}$	nth control site sampled in time step t	Sim.	Randomised: $(SC_{1,t},\ldots,SC_{nc,t}) \sim N(\mu_{C,t}, \sigma_{C,t})$ Non-randomised: $(SC_{1,t},\ldots,SC_{nc,t}) \sim N(\mu_{C,t} + d_{CIB}, \sigma_{C,t})$	-	-	-	-
n_I	Number of impact sites sampled	Sim.	$n_I = \{1,5,10,25,50\}$	-	-	1	50
n_C	Number of control sites sampled	Sim.	$n_C = \{1,5,10,25,50\}$	-	-	1	50

To account for non-random allocation of sites to control and impact groups in BACI, BA, CI and After designs, we repeated the same approach but with one important modification. We adjusted the true density of control sites in every time step, $\mu_{C,t}$, by an empirically derived amount, d_{CIB} ($\mu_{C,t}$+ d_{CIB}; Fig.1; Table 2). To vary d_{CIB}, we used empirical estimates of the proportional difference between control and impact sites in the before period, p_{CIB}, sampled from 47 ecological datasets (d_{CIB}= λ ·(p_{CIB}-1); Table 2; Appendix S1). This simulated difference between control and impact sites accounted for different levels of site selection bias in non-randomised designs, including situations where little or no bias may be present (e.g., d_{CIB}≈0).

We calculated effect size estimates for each design by first finding the mean density of sampled sites across all time steps for control and impact groups in the before period ($Before_{Impact}$, $Before_{Control}$) and the after period ($After_{Impact}$, $After_{Control}$). We assumed that sampling occurred in all time steps ($n_T = T$) in both periods. We did this as the investigator may wish to only estimate the effect over a certain timescale (which will be context-specific) and we lacked the computational capabilities to simulate all possible sampling permutations using fewer than the full number of time steps (e.g., sampling in certain intervals or continuous periods of time; Wauchope et al., 2019).

Effect size estimates were calculated using these mean densities, as appropriate for each study design (Table 3). For example, RCI effect sizes were found by subtracting $After_{Control}$ from $After_{Impact}$, whilst BA effect sizes were found by subtracting $Before_{Impact}$ from $After_{Impact}$ (Table 3). The exception was the After design, for which we found the mean of sampled sites in the first time step and subtracted this from the mean of sampled sites in the final time step of the after period (Table 3). We defined the true effect (Table 3) as the true change in impact sites between the before and after periods (I; Table 2) minus the true change in control sites between the before and after periods (C; Table 2).

We ran the simulation under 1,000 different scenarios, varying: (i) the true change in control sites (C); (ii) the true change in impact sites (I); (iii) the mean difference between control and impact sites in the before period (d_{CIB}); and (iv) the variation between sites (CV_S). For each simulation scenario, we varied the number of time steps simulated (T = 2, 4, 6, 8 or 10), as well as the number of impact sites (n_I = 1, 5, 10, 25, 50) and control sites (n_C= 2, 5, 10, 25, 50) sampled independently to use every possible pairwise combination – a total of 125 combinations. Overall, we simulated 1,000 scenarios with 125 different sampling combinations in each, repeating each scenario 1,000 times (1,000 x (1,000 x 125) = 1.25 x 10^8 runs).

Table 3 *– Equations showing effect size estimate, variance and error calculation for each study design using mean densities of control or impact sites in each period (e.g., $After_{Impact}$ refers to the mean of sampled impact sites across all time steps in the after period). For the After design, the effect size was calculated by finding the difference between the final time step (t=T) and the first time step of the after period (1). n and s^2 refer to the number of sites and variance in that period (e.g., n_{AI} and s^2_{AI} refer to the number of impact sites and variance in the After period). I is the true change in impact sites between the before and after periods, and C is the true change in control sites between the before and after periods (Table 2).*

Design	Effect size estimate	Pooled variance (s_p^2)	Error
RCI	$After_{Impact} - After_{Control}$	$\frac{\left((n_{AI}-1)s^2_{AI} + (n_{AC}-1)s^2_{AC}\right)}{n_{AI}+n_{AC}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}} + \frac{s_p^2}{n_{AC}}}$
BACI	$\left(After_{Impact} - After_{Control}\right) - \left(Before_{Impact} - Before_{Control}\right)$	$\frac{\left((n_{BI}-1)s^2_{BI} + (n_{BC}-1)s^2_{BC} + (n_{AI}-1)s^2_{AI} + (n_{AC}-1)s^2_{AC}\right)}{n_{BI}+n_{BC}+n_{AI}+n_{AC}-4}$	$\pm 1.96 \cdot \sqrt{\left(\frac{s_p^2}{n_{BI}} + \frac{s_p^2}{n_{BC}} + \frac{s_p^2}{n_{AI}} + \frac{s_p^2}{n_{AC}}\right)}$
CI	$After_{Impact} - After_{Control}$	$\frac{\left((n_{AI}-1)s^2_{AI} + (n_{AC}-1)s^2_{AC}\right)}{n_{AI}+n_{AC}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}} + \frac{s_p^2}{n_{AC}}}$
BA	$After_{Impact} - Before_{Impact}$	$\frac{\left((n_{BI}-1)s^2_{BI} + (n_{AI}-1)s^2_{AI}\right)}{n_{BI}+n_{AI}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{AI}} + \frac{s_p^2}{n_{BI}}}$
After	$After_{Impact,t=T} - After_{Impact,t=1}$	$\frac{\left((n_{I,t=T}-1)s^2_{I,t=T} + (n_{I,t=1}-1)s^2_{I,t=1}\right)}{n_{I,t=T}+n_{I,t=1}-2}$	$\pm 1.96 \cdot \sqrt{\frac{s_p^2}{n_{I,t=T}} + \frac{s_p^2}{n_{I,t=1}}}$
True effect	$I - C$		

Effect size estimates for each design were used to investigate their relative accuracy. We calculated 95% Confidence Intervals using the pooled variance and associated error for each effect size estimate (Table 3). We used these 95% Confidence Intervals to estimate the percentage of simulations repetitions where: (i) the true effect fell within the 95% Confidence Intervals of effect size estimates (coverage probability); (ii) the correct direction was detected (95% Confidence Intervals entirely above or below zero); (iii) the estimated effect size under- or overestimated the true effect (95% Confidence Intervals entirely above or below true effect). We also investigated the percentage of simulation repetitions in which each design's effect size estimate: (i) had the same absolute direction as the true effect; and (ii) was both within a given percentage of the true effect *and* of the same direction (based on 95% Confidence Intervals). We believe the latter measure captures the major aspects of accuracy and precision that are desirable in a study design.

We calculated all these measures for all possible pairwise combinations of control and impact sites (e.g., two control and two impact sites, two control and five impact sites, etc.). We set five thresholds to measure the percentage of times an effect size estimate was within a certain percentage of the true effect: 10, 20, 30, 40 and 50%. We also explored how varying the terms *C* and d_{CIB} (affecting bias in BA, CI, and After designs) for three levels of magnitude (no bias: 1; low bias: 0.9 or 1.1; high bias: 0.7 or 1.3) affected this percentage (Figures S7 and S8).
We used Generalised Linear Models (GLMs) with a beta error distribution to determine the relationship between the performance of each design (the response variable; see below) and two explanatory variables (number of control sites and the number of impact sites). A beta GLM was most appropriate as it provides a flexible method to analyse proportional data bounded between 0 and 1 and accounts for heteroskedasticity and asymmetric distributions. For the response variable, we used the proportion of simulation repetitions where the effect size estimate was within ±30% of the true effect and had the correct direction. We only considered results for an accuracy threshold of ±30% as this was deemed a reasonable level of accuracy and we wanted to simplify the interpretation of our results as much as possible. We also present results for other accuracy thresholds (±10% and ±50%) in Appendix S3.

Based on graphical observations of the relationship between the response and explanatory variables, we included impact and control sites as log transformed explanatory variables for models of RCI and BACI designs and tested models with and without an interaction term between these variables (Appendix S4). BACI and RCI models with both impact sites and control sites as predictor variables without an interaction term were selected as the best models because they had the lowest values of AIC; although the model using an interaction term between impact and control sites was within 2 units of AIC, we selected the models without an interaction term because they were more parsimonious (Appendix S4). As the performance of BA, CI and After designs did not vary with the number of impact or control sites, we did not create any models for these designs. We calculated quasi-R^2 values (Appendix S4) to test model performance using the equation:

$$quasiR^2 = 1 - \frac{deviance}{null\ deviance} \quad \text{(Equation 1).}$$

Both of the selected models were only slightly over-dispersed (RCI model: θ = 1.19; BACI model: θ = 1.19) and Pearson's χ2 residuals were non-significant ($p>0.05$) suggesting no significant patterns remained in the residuals. There were also no observable patterns between residuals and explanatory variables or fitted values.

For RCI and BACI designs, we converted estimated coefficients (β) from log odds (Appendix S4) to proportions to create an 'accuracy weight' equation for each design (Equations 2 and 3). We found the accuracy weights for BA, CI, and After designs by simply taking the mean

proportion of simulation repetitions where the effect size estimate was within ±30% of the true effect and had the correct direction (based on 95% Confidence Intervals) across all combinations of impact and control sites. These accuracy weights are on a continuous scale between a minimum of 0 (lowest accuracy) and a maximum of 1 (highest accuracy – see Results, Discussion and Appendix S2 for how to apply these weights):

$$\text{RCI accuracy weight} = \frac{1}{1+e^{-\left(\beta_{RCI\_Int.}+\beta_{RCI\_n_I}\cdot\ln(n_I)+\beta_{RCI\_n_C}\cdot\ln(n_C)\right)}} \quad \text{(Equation 2)}$$

$$\text{BACI accuracy weight} = \frac{1}{1+e^{-\left(\beta_{BACI\_Int.}+\beta_{BACI\_n_I}\cdot\ln(n_I)+\beta_{BACI\_n_C}\cdot\ln(n_C)\right)}} \quad \text{(Equation 3)}$$

where $\beta_{BACI\_Int.}$ = BACI model intercept coefficient, $\beta_{BACI\_n_I}$= BACI model impact sites coefficient, $\beta_{BACI\_n_C}$= BACI model control sites coefficient ($\beta_{RCI\_Int.}$, $\beta_{RCI\_n_I}$, and $\beta_{RCI\_n_C}$ refer to the equivalent coefficients for the RCI model).

We applied our accuracy weights to three recent ecological meta-analyses: Bernes et al. (2018) on the effects of ungulate herbivory on vegetation and invertebrates; Eales et al. (2018) on the effects of prescribed burning on forest biodiversity; and Sandström et al. (2019) on the impacts of dead wood manipulation on forest biodiversity. We found these meta-analyses by searching the Environmental Evidence Journal and the Journal of Applied Ecology using the search terms: “meta analysis” OR “meta-analysis” and reviewing studies published since 2018. Only the three previously mentioned meta-analyses contained a sufficient range of study designs (Appendix S2) and readily available associated data on study design, replicates, and effect sizes. We repeated analyses using random effects models following the authors’ methodology (e.g., including random factors such as Site IDs) using the metafor package (Viechtbauer, 2010); see Appendix S2). We were able to replicate 128 out of 130 summary effect sizes (comparisons) using the authors’ methodology and inverse-variance weighting, which we repeated with our accuracy weights. Two summary effect sizes could not be replicated from Bernes et al. (2018) due to lack of data labelling. We wanted to test how our weights altered the conclusions of meta-analyses that used studies with a mixture of different study designs. Therefore, we only present results for 96 comparisons that used studies with at least one type of design. The mean number of studies of each design were: 9.0 BACI, 6.0 BA, 5.0 CI (see Appendix S2 for a breakdown of studies for each summary effect size).

We used R statistical software version 3.5.1 (R Core Team, 2019) with the doParallel package (Microsoft Corporation and Weston, 2019) to increase computational performance. Simulation R code and empirical data for simulations (see Appendix S5 for published data used in simulations) are available via Zenodo: https://doi.org/10.5281/zenodo.4437010.

Results

There was large variation in the performance of designs in accurately estimating the true effect. As overall patterns were similar across simulations with different time steps (Figures S1-S3), we present results when six time steps were simulated in both the before and after periods.

RCI designs performed best at correctly identifying the direction of the true effect (≥91.2% of simulation repetitions; Fig.2A), followed by BACI (≥87.4% of the time). Both BACI and RCI far outperformed CI, BA, and particularly After designs – CI designs performed similarly to BA designs (approximately 75.2% versus 73.9%, respectively) and both strongly outperformed After designs (approximately 49.5%; Fig.2A). All designs showed negligible improvements in performance with increasing replication, particularly BA, CI, and After designs (increases from two control and two impact sites to 50 control and 50 impact sites: RCI = +0.9%; BACI = +1.5%; BA = +0.5%; CI= +0.2%; After = 0.3%; Fig.2A).

Taking account of the uncertainty around these effect size estimates (95% Confidence Intervals) gave broadly similar results – where overlap with zero was classed as non-significant and non-overlap as either positive or negative (Fig.2B). With this measure, RCI was most likely to correctly predict the direction of the true effect with two impact and control sites, followed by BACI, CI, BA, and After designs in decreasing order of performance (Fig.2B). BACI designs showed the greatest proportional improvement in this measure, performing worse than CI designs at a sample size of 2 impact and control sites, but performing substantially better at much larger sample sizes (Fig.2B). The accuracy of BA, CI, and After designs in significantly identifying the correct direction did improve noticeably with sample size, but less so compared to RCI and BACI designs (Fig.2B). RCI and BACI designs were much less likely to produce significant effect sizes that had the wrong direction (1-7% and 1-5% of repetitions, respectively, depending on the sample size; Figure S3).

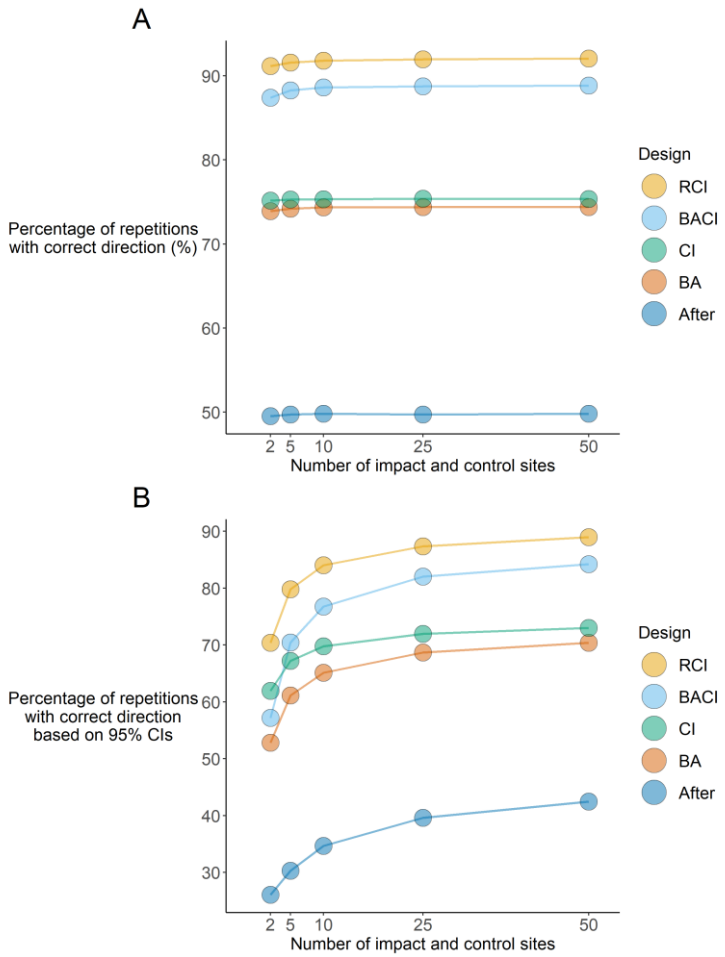


Figure 2 *– Performance of designs in correctly predicting the direction of the true effect for multiple levels of spatial replication with equal numbers of control and impact sites (see Figures S1 and S2 for other combinations of sites). Fig.2A measures this in terms of whether the effect size estimate was positive or negative, whilst Fig.2B considers whether the 95% Confidence Intervals of this estimate correctly fell entirely above or below zero. See Table 1 for the definition of each design.*

If we consider the coverage probabilities of each design (i.e., proportion of times the true effect fell within the 95% Confidence Intervals of effect size estimates), RCI and BACI designs substantially outperformed BA, CI, and After designs (Fig.3A). The coverage probabilities for all designs declined asymptotically with increasing sample size as 95% Confidence Intervals narrowed (Fig.3A).

We also examined the tendency for designs to underestimate or overestimate the true effect (i.e., when 95% Confidence Intervals did not overlap with true effect; Fig.3B). RCI and BACI designs were substantially less likely to underestimate the true effect compared BA, CI, and After designs and less likely to overestimate the true effect compared After designs (Fig.3B). RCI and BACI designs were approximately as likely to overestimate the true effect as BA and CI designs (slightly less likely at lower sample sizes and slightly more likely at higher sample

sizes) and were as likely to overestimate as to underestimate. After designs were also approximately as likely to underestimate as to overestimate, but BA and CI designs were substantially more likely to underestimate than overestimate (Fig.3B). All designs were increasingly likely to underestimate or overestimate the true effect with increasing sample size as 95% Confidence Intervals narrowed (in line with decreases in coverage probability; Fig.3A).

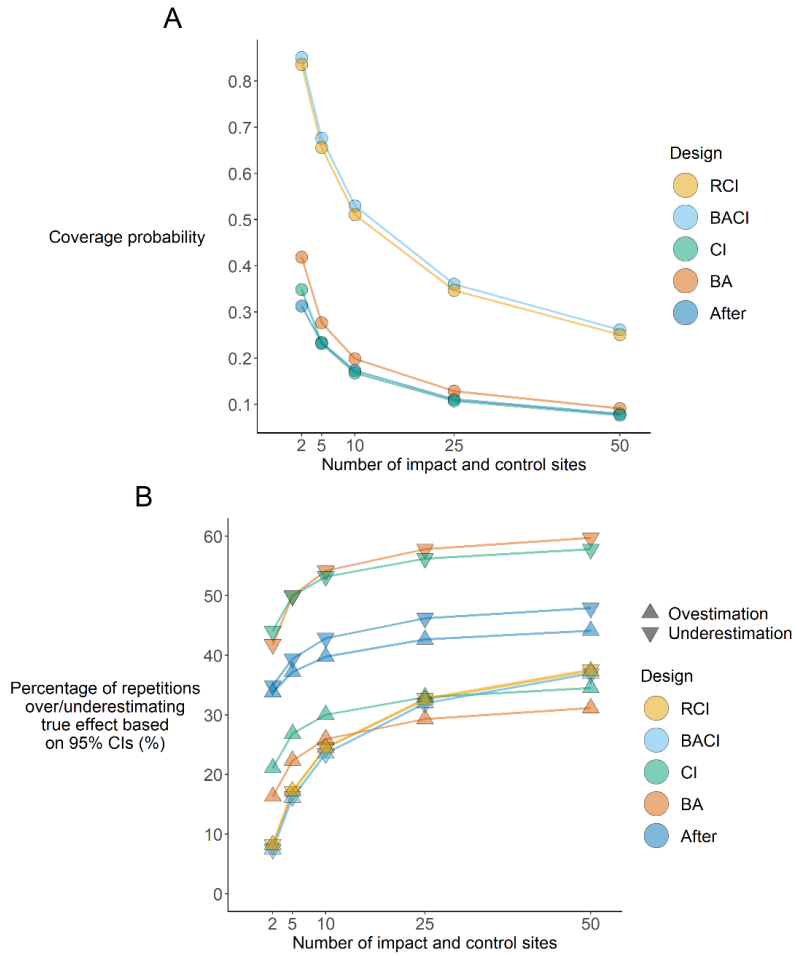


Figure 3 *– Percentage of simulation repetitions in which the 95% Confidence Intervals of effect size estimates contained the true effect (coverage probability – Fig.3A) or were either greater than or less than the true effect (overestimate versus underestimate – Fig.3B). In Fig.3B, underestimates are shown by downward triangles, whilst overestimates are shown by upward triangles. This is shown for multiple levels of spatial replication with equal numbers of control and impact sites (see Figures S4 and S5 for other combinations of sites). See Table 1 for the definition of each design.*

Consistent patterns were found when considering the percentage of repetitions for which the effect size estimate was both within a certain percentage of the true effect and had the correct direction based on 95% Confidence Intervals (Fig.4). First, RCI and BACI designs far outperformed BA, CI, and After designs (for ±30% accuracy threshold: RCI ≥46.7%, BACI ≥31.1%, BA ≥19.0%, CI ≥17.6%, After ≥6.4%; Fig.4). Second, BA designs appeared to

perform slightly better than CI designs, but only noticeably as the accuracy threshold rose above 30% (Fig.4). Similarly, both BA and CI designs performed relatively better compared to After designs with an increasing accuracy threshold (Fig.4).

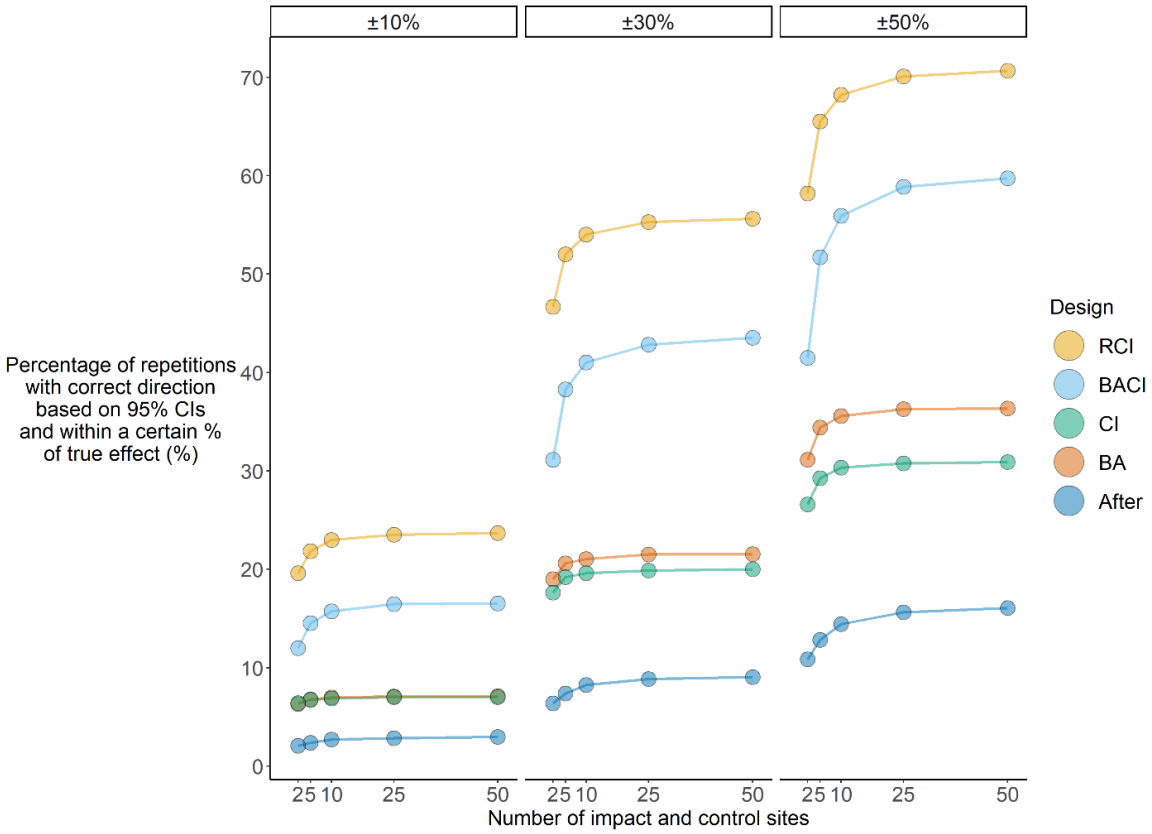


Figure 4 *– Performance of designs measured by percentage of simulation repetitions in which a design's effect size estimate was both within ±10, 30, or 50% of the true effect and had the correct direction based on 95% Confidence Intervals. This is shown for multiple levels of spatial replication with equal numbers of control and impact sites (see Figure S6 for other combinations of sites). See Table 1 for the definition of each design.*

Third, RCI and BACI performance increased to a greater extent with increasing replication than for other designs (Fig.4). For the ±30% accuracy threshold, increasing replication from two control and impact sites to 50 control and impact sites resulted in an increase of +9.0% for RCI and 12.4% for BACI compared to +2.5% for BA, +2.4% for CI and 2.7% for After (Fig.4). For RCI and BACI designs, increasing replication moderately in both control and impact sites resulted in greater performance than only increasing replication in just one type of site (±30% threshold: 52.0% and 38.3%, respectively at five impact and five control sites versus 51.0% and 36.8%, respectively, at two impact and 50 control sites; Fig.S6).

We also considered how varying the simulation parameters C and d_{CIB} affected our results (Figures S7 and S8). Increasing the change in control (C) reduced the performance of BA

designs substantially (Figure S7), whilst increasing the initial mean differences between impact and control groups in the before period (d_{CIB}) reduced the performance of CI designs substantially (Figure S8).

We used Generalised Linear Models (GLMs) to examine how the sample size of impact and control sites determined the performance of each design at estimating the direction and magnitude of the true effect to within ±30% (using data from Fig.4 and Fig.S6). For RCI and BACI designs, there was little difference in the importance of control versus impact sites in predicting performance. High Pseudo-R^2 values showed that our models explained far greater levels of variation in the data than null models (see Appendix S4).

Weights for studies in meta-analyses can be calculated from these relationships of performance with sample size, which we term 'accuracy weights'. This requires information about a study's design and the number of independent control and impact units used (see Appendix S4). For example, França et al. (2016) used a BACI design, 29 impact and five control units and thus receives an accuracy weight of:

$$\frac{1}{1 + e^{-(-0.813 + 0.0792 \cdot \ln(29) + 0.0797 \cdot \ln(5))}} = 0.397$$

(see Appendix S4; Equation 2). See Appendix S2 for more examples of study weights.

We applied our accuracy weights to meta-analyses (Appendix S2) by multiplying the conventional inverse-variance weight matrix by a matrix of our accuracy weights. We found that they gave broadly similar results to conventional inverse-variance weighting (for 94% of comparisons). However, there was a small tendency for our weights to alter the outcome to non-significant (3% from negative to non-significant and 3% from positive to non-significant; Table 4). No outcomes of summary effect sizes changed from positive to negative or vice versa, or from non-significant to significantly positive (Table 4).

Table 4 *– Comparison of outcomes for 96 summary effect sizes obtained using the accuracy weights proposed by this study versus conventional inverse-variance weighting. Summary effect sizes were extracted from 3 separate meta-analyses (see Materials and methods). Cells show the proportion of effect sizes that were significantly positive, significantly negative, or non-significant for both weighting systems.*

Weighting method		**Accuracy weight**		
	Outcome	+	Non-significant	-
Inverse-variance weight	+	14 (15%)	3 (3%)	0
	Non-significant	0	56 (58%)	0
	-	0	3 (3%)	20 (21%)

Discussion

Using this simulation, we have demonstrated that RCI and BACI designs are far more accurate than BA, CI and After designs. When estimating the true effect to within ±30% and correctly identifying its direction, RCI performed, depending on sample size, 2.5-2.6 times better than BA, 2.6-2.8 times than CI, and 6.1-7.3 times than After designs, whilst BACI performed 1.6-2.0 times better than BA, 1.8-2.2 times than CI, and 4.8-5.2 times than After designs. RCI performed approximately 1.1-1.3 times better than BACI. Increasing sample size tends to only increase precision in simpler designs (CI, BA, and After) around a more biased estimate of the true effect.

This bias is generated by violating the assumptions underpinning these simpler designs (de Palma et al., 2018). BA designs assume there is no average change in the control group mean before versus after the intervention, whilst CI designs assume the only differences that exist between control and impact sites are due to the focal impact (C and d_{CIB} in Materials and methods; Figures S7 and S8). After designs make both these assumptions and therefore can only claim to provide information on the rate of change in a measured variable over time following an impact (de Palma et al., 2018). Blocking, pairing, or matching sites in CI designs or including a proxy variable in statistical analysis could theoretically account for some of this bias, but they cannot guarantee lower levels of bias. CI designs (often called Space-for-Time substitutions when quantifying land-use change) also suffer from additional biases that we did not include in our simulation (e.g., biotic lag; see de Palma et al., 2018) and so our quantification of the inaccuracy of this design is likely to be an underestimate.

RCI does not suffer from bias, but rather statistical noise, because the use of randomisation effectively eliminates confounding biases; initial differences generated by spatiotemporal variation become stochastic noise that can be minimised using larger sample sizes (de Palma et al., 2018). RCI and BACI designs better account for pre-impact differences between impact and control groups, either through randomisation or sampling impact and control groups before and after the impact, respectively. RCI and BACI designs therefore have higher accuracy and their performance can be improved through repeated sampling through time to better account for spatiotemporal variation (Thiault et al., 2017). BACI designs however, unlike RCI designs, are not randomised and so still suffer from some bias because of the 'parallel trends' assumption (Angrist and Pischke, 2008; Ding and Li, 2019). However, BACI designs still performed substantially better than the simpler non-randomised designs.

The fact that increasing the sample size (precision) of simpler designs reduced the coverage probability (probability that the true effect fell within the 95% Confidence Intervals of an effect size estimate; Fig.3A) supports this conclusion as 95% Confidence Intervals converged on

biased estimates. The coverage probability for RCI and BACI designs remained higher because the effect size estimates tend to converge on the true effect – greater precision was more likely to translate into greater accuracy. Greater accuracy is therefore best achieved using more robust designs that remove biases, such as RCI and BACI designs.

We nevertheless note that of the two better performing designs, BACI designs in particular are known to suffer from noise (Osenberg et al., 2006), which was demonstrated by their difficulty in correctly predicting the direction of the true effect in terms of 95% Confidence Intervals (significance) at low samples sizes (Fig.2B). The low statistical power of BACI designs at small sample sizes is an issue and reinforces the need to ensure sufficient numbers of replicates are used in BACI designs (Osenberg et al., 2006), as well as to consider using Bayesian approaches to interpret effect sizes (Conner et al., 2016). At larger sample sizes however, RCI and BACI designs predicted the (statistical significant) direction of the true effect correctly far more frequently than CI and BA designs, and even more so than After designs (Fig.2B).

Our results provide strong evidence that simpler designs (e.g., After, BA, and CI) often yield different inferences to RCI and BACI designs, as observed empirically by previous studies (França et al., 2016; Mahlum et al., 2018; Osenberg et al., 2011; Smokorowski et al., 2017). We also found that BA and CI designs were more prone to underestimation than overestimation (Fig.3B), which is consistent with results from França et al. (2016) that showed a CI design underestimated the impacts of logging relative to a BACI design. Therefore, we argue that studies using After, BA and CI designs risk presenting misleading conclusions on the impact of threats and interventions. To our knowledge this simulation is not only the first quantitative comparison to demonstrate this, but also to show *how* inaccurate these simpler designs may typically be in ecology under varying levels of spatial replication.

We have confidence in these conclusions as we used empirically derived parameter estimates from 47 ecological datasets to quantify the likelihood and magnitude of the biases that affect study designs in ecology (d_{CIB} and C; Materials and methods; Appendix S1). The context-dependency of our results, linked to how the likelihood and magnitude of biases varies across different fields of ecology, could be investigated using our R code if sufficient empirical data is available to characterise major parameters (d_{CIB} and C; Figures S7 and S8) in different contexts. Future work could explore the effects of different types of trends and lag periods on the relative performance of designs, since previous literature has often assumed there is no overall pre-impact trend – only fluctuations around a baseline average (Lettenmaier et al., 1978; Thiault et al., 2017). Nevertheless, our results provide strong evidence that, generally in ecology, we should invest in implementing more robust designs whenever possible – investing effort into using simpler designs with greater sample sizes is simply inefficient.

Although we strongly advocate for greater investment in more robust designs, we also realise there is a trade-off between the greater accuracy of robust designs and greater logistical ease of simpler designs. Whilst we can generate more studies with simpler designs more easily, their probable low accuracy means that we may use misleading evidence to inform policy and practice. We nevertheless argue that situations still arise where investigators can use robust designs and yet fail to; promoting greater awareness of more robust designs and opportunities for their usage is important. For example, BACI design use should be encouraged whenever prior knowledge exists of the timing of an impact or where suitable pre-impact data is available retrospectively (e.g., infrastructure projects, Protected Area designation).

We also recognise the expensive nature of BACI designs, due to the need to revisit study sites before and after the impact, often hampers their implementation (de Palma et al., 2018). This means that BACI designs can be challenging to use during short term projects limited by grant or studentship duration. However, we would also argue that the costs of misinforming decision-makers and making inaccurate inferences (including both the costs to the credibility of scientific evidence and the costs involved in implementing an ineffective intervention) may outweigh the costs saved in conducting simpler study designs (Wauchope, 2020, p127-128). For example, delaying decisions until a more expensive, but more rigorous study design can be implemented has been shown to be an optimal strategy, even in a crisis discipline such as biodiversity conservation (Iacona et al. 2017). Any cost-based assessment of the feasibility of a particular study design should incorporate the socio-political and environmental costs of Type I and Type II errors associated with the proposed design (Mapstone, 1995). For example, important interventions or impacts that carry greater risk should warrant the implementation of a higher minimum standard of study design (Mapstone, 1995). Researchers should adjust budgets and project plans to accommodate study designs, rather than the other way around.

Therefore, we suggest that longer-term funding and stronger research-practice partnerships are urgently required to facilitate the use of RCI and BACI designs (de Palma et al., 2018; Osenberg et al., 2011). Alongside greater promotion of more rigorous designs, it would also be helpful to promote approaches that aid the ecological interpretation of results, such as Bayesian methods to generate more easily interpretable probabilities for managers and practitioners (Conner et al., 2016) and partitioning BACI design results into two measures (CI-divergence and CI-contribution; see Chevalier et al., 2019).

Given the use of simpler designs will probably persist in the near future, we further argue that our results have major implications for decision-making and meta-analysis in ecology. We have proposed a novel weighting system that could help when meta-analyses are faced with studies that vary markedly in their design. Conventional meta-analyses typically use inverse-

variance of studies as weights to attempt to account for study quality (Koricheva and Gurevitch, 2014; Marín-Martínez and Sánchez-Meca, 2010). However, this can greatly reduce the number of suitable primary studies since not all studies report variance (Koricheva and Gurevitch, 2014). Alternative approaches of meta-analysis to tackle poor data reporting, such as non-parametric weighting by sample size, have been proposed (Adams et al., 1997; Mayerhofer et al., 2013), but fail to consider wider aspects of study quality such as study design (Spake and Doncaster, 2017). Whilst recent efforts to assessing evidence quantitatively by study design are welcomed (Mupepele and Dormann, 2016; Webb et al., 2012), their weights are relatively simplistic (e.g., simple integer scores or categories) and lack a quantitative or objective grounding.

Our weighting system is informed by the relationships we have found between accuracy, precision, study design, and sample size, arguably accounting for more aspects of study quality than weighting by sample size or inverse-variance. We have shown how our accuracy weights can be applied to meta-analyses to give greater influence to studies with more accurate designs, in addition to weighting by inverse-variance (Appendix S2). We have also demonstrated that, although there was generally good agreement between these systems of weighting, our accuracy weighting approach tended to reduce the number of positive and negative significant results in meta-analyses (Table 4). This suggests that inverse-variance weighting may have led to more significantly positive or negative results by erroneously rewarding studies with simpler designs when they have higher precision (lower variance). This is problematic because we have shown that increasing the precision of simpler designs does not improve accuracy and often leads to biased estimates. Weighting by a combination of accuracy and precision using our weighting approach seems more sensible given these results.

Although we acknowledge that our weights only consider some aspects of study quality, we believe that they could be modulated using the percentage of criteria met in subject-specific quality checklists to incorporate more context-specific factors (e.g., size of sampling unit, temporal replication, and internal validity; Bilotta et al., 2014; Mupepele and Dormann, 2016). Adding extra components to the evidence assessment process, however, must be balanced against the effort expended in doing so. This is particularly important as the growth of scientific evidence bases accelerates (Bornmann and Mutz, 2015; Larsen and von Ins, 2010) because we will need to design systems of assessing and critically appraising evidence that are faster and more efficient to ensure evidence synthesis can keep pace (Marshall and Wallace, 2019; O'Connor et al., 2018; Thomas et al., 2017; Wallace et al., 2014). Automation of evidence assessment using a weighting system such as ours is one potential solution that could speed up evidence synthesis (Marshall et al., 2020, 2015; Marshall and Wallace, 2019; O'Connor et

al., 2018; Tsafnat et al., 2013; Wallace et al., 2014), but of course this needs to be rigorously tested and developed further to ensure it gives a reliable reflection of the quality of a study. At a coarser scale, our weights could also assign studies to different accuracy categories (Appendix S2), giving a rapid, easily interpretable way to communicate the robustness of evidence to decision-makers – e.g., in evidence toolkits such as Conservation Evidence (Sutherland et al., 2019). We welcome future research to explore how best to apply, develop, and strengthen our accuracy weighting approach within evidence assessment and decision-making processes.

Overall, we have shown for the first time how much less accurate simpler study designs are compared to more complex ones, generating a new quantitative understanding of the relative accuracy of different designs. Further refinement and development of our accuracy weighting approach could also offer a powerful, yet versatile new approach to weighting evidence where studies use a range of different designs, with major implications for the future of meta-analysis and decision-making in automating evidence assessment. We hope our work encourages greater discussion of study design by scientists, managers, and policymakers across ecology and demonstrates the need to tackle the serious consequences of using different designs to make inferences in ecology.

References

Adams, D.C., Gurevitch, J., Rosenberg, M.S., 1997. Resampling tests for meta-analysis of ecological data. Ecology 78, 1277–1283. https://doi.org/10.1890/0012-9658

Angrist, J.D., Pischke, J.-S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, New Jersey.

Bernes, C., Bullock, J.M., Jakobsson, S., Rundlöf, M., Verheyen, K., Lindborg, R., 2017. How are biodiversity and dispersal of species affected by the management of roadsides? A systematic map. Environmental Evidence 6, 24. https://doi.org/10.1186/s13750-017-0103-1

Bernes, C., Jonsson, B.G., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., Sandström, J., 2015. What is the impact of active management on biodiversity in boreal and temperate forests set aside for conservation or restoration? A systematic map. Environmental Evidence 4, 25. https://doi.org/10.1186/s13750-015-0050-7

Bernes, C., Macura, B., Jonsson, B.G., Junninen, K., Müller, J., Sandström, J., Lõhmus, A., Macdonald, E., 2018. Manipulating ungulate herbivory in temperate and boreal forests: effects on vegetation and invertebrates. A systematic review. Environmental Evidence 7, 13. https://doi.org/10.1186/s13750-018-0125-3

Bilotta, G.S., Milner, A.M., Boyd, I.L., 2014. Quality assessment tools for evidence from environmental science. Environmental Evidence 3, 14. https://doi.org/10.1186/2047-2382-3-14

Bornmann, L., Mutz, R., 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66, 2215–2222. https://doi.org/https://doi.org/10.1002/asi.23329

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70–79. https://doi.org/10.1080/01621459.1975.10480264

Burge, O.R., Bodmin, K.A., Clarkson, B.R., Bartlam, S., Watts, C.H., Tanner, C.C., 2017. Glyphosate redirects wetland vegetation trajectory following willow invasion. Applied Vegetation Science 20, 620–630. https://doi.org/10.1111/avsc.12320

Chevalier, M., Russell, J.C., Knape, J., 2019. New measures for evaluation of environmental perturbations using Before-After-Control-Impact analyses. Ecological Applications 29, e01838. https://doi.org/10.1002/eap.1838

Conner, M.M., Saunders, W.C., Bouwes, N., Jordan, C., 2016. Evaluating impacts using a BACI design, ratios, and a Bayesian approach with a focus on restoration. Environmental Monitoring and Assessment 188, 555. https://doi.org/10.1007/s10661-016-5526-6

Damgaard, C., 2019. A Critique of the Space-for-Time Substitution Practice in Community Ecology. Trends in Ecology and Evolution 34, 416–421. https://doi.org/10.1016/j.tree.2019.01.013

de Palma, A., Sanchez-Ortiz, K., Martin, P.A., Chadwick, A., Gilbert, G., Bates, A.E., Börger, L., Contu, S., Hill, S.L.L., Purvis, A., 2018. Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity: Study Design, Time, Space and Synthesis. Next Generation Biomonitoring 58, 163–199. https://doi.org/https://doi.org/10.1016/bs.aecr.2017.12.004

di Fonzo, M., Collen, B., Mace, G.M., 2013. A new method for identifying rapid decline dynamics in wild vertebrate populations. Ecology and Evolution 3, 2378–2391. https://doi.org/10.1002/ece3.596

Dietl, G.P., Durham, S.R., 2016. Data from: Geohistorical records indicate no impact of the Deepwater Horizon oil spill on oyster body size. Dryad Digital Repository. https://doi.org/10.5061/dryad.bc80t

Ding, P., Li, F., 2019. A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. Political Analysis 27, 605–615. https://doi.org/DOI: 10.1017/pan.2019.25

Downs, S.H., Black, N., 1998. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. Journal of Epidemiology and Community Health 52, 377 LP – 384. https://doi.org/10.1136/jech.52.6.377

Eales, J., Haddaway, N.R., Bernes, C., Cooke, S.J., Jonsson, B.G., Kouki, J., Petrokofsky, G., Taylor, J.J., 2018. What is the effect of prescribed burning in temperate and boreal forest on biodiversity, beyond pyrophilous and saproxylic species? A systematic review. Environmental Evidence 7, 19. https://doi.org/10.1186/s13750-018-0131-5

Fernández-Chacón, A., Moland, E., Espeland, S.H., Olsen, E.M., 2015. Demographic effects of full vs. partial protection from harvesting: inference from an empirical before–after control-impact study on Atlantic cod. Journal of Applied Ecology 52, 1206–1215. https://doi.org/10.1111/1365-2664.12477

França, F., Louzada, J., Korasaki, V., Griffiths, H., Silveira, J.M., Barlow, J., 2016. Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. Journal of Applied Ecology 53, 1098–1105. https://doi.org/10.1111/1365-2664.12657

Hewitt, J.E., Thrush, S.E., Cummings, V.J., 2001. Assessing environmental impacts: effects of spatial and temporal variability at likely impact scales. Ecological Applications 11, 1502–1516. https://doi.org/https://doi.org/10.1890/1051-0761(2001)011[1502:AEIEOS]2.0.CO;2

Hipel, K.W., Lettenmaier, D.P., McLeod, A.I., 1978. Assessment of environmental impacts part one: Intervention analysis. Environmental Management 2, 529–535. https://doi.org/10.1007/BF01866711

Huusela-Veistola, E., 1998. Effects of perennial grass strips on spiders (Araneae) in cereal fields and impact on pesticide side-effects. Journal of Applied Entomology 122, 575–583. https://doi.org/10.1111/j.1439-0418.1998.tb01548.x

Iacona, G.D., Possingham, H.P., Bode, M., 2017. Waiting can be an optimal conservation strategy, even in a crisis discipline. Proceedings of the National Academy of Sciences 114, 10497 LP – 10502. https://doi.org/10.1073/pnas.1702111114

Koricheva, J., Gurevitch, J., 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology 102, 828–844. https://doi.org/https://doi.org/10.1111/1365-2745.12224

Larsen, A.E., Meng, K., Kendall, B.E., 2019. Causal analysis in control–impact ecological studies with observational data. Methods in Ecology and Evolution 10, 924–934. https://doi.org/10.1111/2041-210X.13190

Larsen, P.O., von Ins, M., 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84, 575–603. https://doi.org/10.1007/s11192-010-0202-z

Lettenmaier, D.P., Hipel, K.W., McLeod, A.I., 1978. Assessment of environmental impacts part two: Data collection. Environmental Management 2, 537–554. https://doi.org/10.1007/BF01866712

Mahlum, S., Cote, D., Wiersma, Y.F., Pennell, C., Adams, B., 2018. Does restoration work? It depends on how we measure success. Restoration Ecology 26, 952–963. https://doi.org/10.1111/rec.12649

Mapstone, B.D., 1995. Scalable Decision Rules for Environmental Impact Studies: Effect Size, Type I, and Type II Errors. Ecological Applications 5, 401–410. https://doi.org/10.2307/1942031

Marín-Martínez, F., Sánchez-Meca, J., 2010. Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. Educational and Psychological Measurement 70, 56–73. https://doi.org/10.1177/0013164409344534

Marshall, I.J., Johnson, B.T., Wang, Z., Rajasekaran, S., Wallace, B.C., 2020. Semi-Automated evidence synthesis in health psychology: current methods and future prospects. Health Psychology Review 14, 145–158. https://doi.org/10.1080/17437199.2020.1716198

Marshall, I.J., Kuiper, J., Wallace, B.C., 2015. Automating Risk of Bias Assessment for Clinical Trials. IEEE Journal of Biomedical and Health Informatics 19, 1406–1412. https://doi.org/10.1109/JBHI.2015.2431314

Marshall, I.J., Wallace, B.C., 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews 8, 163. https://doi.org/10.1186/s13643-019-1074-9

Mayerhofer, M.S., Kernaghan, G., Harper, K.A., 2013. The effects of fungal root endophytes on plant growth: a meta-analysis. Mycorrhiza 23, 119–128. https://doi.org/10.1007/s00572-012-0456-9

Merrow, J., 2007. Effectiveness of Amphibian Mitigation Measures Along a New Highway. UC Davis: Road Ecology Center. https://escholarship.org/uc/item/7bn605dv

Microsoft Corporation, Weston, S., 2019. doParallel: Foreach Parallel Adaptor for the “parallel” Package. R package version 1.0.15.

Moland, E., Olsen, E.M., Knutsen, H., Garrigou, P., Espeland, S.H., Kleiven, A.R., André, C., Knutsen, J.A., 2013. Lobster and cod benefit from small-scale northern marine protected areas: inference from an empirical before–after control-impact study. Proceedings of the Royal Society B: Biological Sciences 280, 20122679. https://doi.org/10.1098/rspb.2012.2679

Mupepele, A.-C., Dormann, C., 2016. Influence of Forest Harvest on Nitrate Concentration in Temperate Streams—A Meta-Analysis. Forests 8, 5. https://doi.org/10.3390/f8010005

O'Connor, A.M., Tsafnat, G., Gilbert, S.B., Thayer, K.A., Wolfe, M.S., 2018. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic Reviews 7, 3. https://doi.org/10.1186/s13643-017-0667-4

Osenberg, C., Bolker, B., White, J.S., St. Mary, C., Shima, J., 2006. Statistical issues and study design in ecological restorations: lessons learned from marine reserves, Foundations of restoration ecology. Island Press, Washington, DC.

Osenberg, C.W., Schmitt, R.J., 1996. Detecting Ecological Impacts Caused by Human Activities, in: Schmitt, R.J., Osenberg, C.W. (Eds.), Detecting Ecological Impacts. Elsevier, San Diego, pp. 3–16. https://doi.org/10.1016/B978-012627255-0/50003-3

Osenberg, C.W., Shima, J.S., Miller, S.L., Stier, A.C., 2011. Assessing effects of marine protected areas: confounding in space and possible solutions, in: Claudet, J. Marine Protected Areas - a Multidisciplinary Approach. Cambridge University Press, New York, pp. 143–167.

Papathanasopoulou, E., Queirós, A.M., Beaumont, N., Hooper, T., Nunes, J., 2016. What evidence exists on the local impacts of energy systems on marine ecosystem services: a systematic map. Environmental Evidence 5, 1–12. https://doi.org/10.1186/s13750-016-0075-6

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Sandström, J., Bernes, C., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., Jonsson, B.G., 2019. Impacts of dead wood manipulation on the biodiversity of temperate and boreal forests. A systematic review. Journal of Applied Ecology 56, 1770–1781. https://doi.org/10.1111/1365-2664.13395

Scoccianti, C., 2006. Rehabilitation of habitat connectivity between two important marsh areas divided by a major road with heavy traffic. Acta Herpetologica 1, 77–79. http://digital.casalini.it/10.1400/16859

Smokorowski, K.E., Randall, R.G., Canada, O., East, Q.S., Canada, O., Lakes, G., 2017. Cautions on using the Before-After-Control-Impact design in environmental effects monitoring programs. Facets 2, 212–232. https://doi.org/10.1139/facets-2016-0058

Spake, R., Doncaster, C.P., 2017. Use of meta-analysis in forest biodiversity research: key challenges and considerations. Forest Ecology and Management 400, 429–437. https://doi.org/10.1016/j.foreco.2017.05.059

Stewart-Oaten, A., Bence, J.R., 2001. Temporal and Spatial Variation in Environmental Impact Assessment. Ecological Monographs 71, 305–339. https://doi.org/10.1890/0012-9615(2001)071[0305:TASVIE]2.0.CO;2

Stewart-Oaten, A., Murdoch, W.W., Parker, K.R., 1986. Environmental Impact Assessment: "Pseudoreplication" in Time? Ecology 67, 929. https://doi.org/10.2307/1939815

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Thiault, L., Kernaléguen, L., Osenberg, C.W., Claudet, J., 2017. Progressive-Change BACIPS: a flexible approach for environmental impact assessment. Methods in Ecology and Evolution 8, 288–296. https://doi.org/10.1111/2041-210X.12655

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, Steven, Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., Albrecht, L., Horsley, T., Platt, J., Armstrong, R., Nguyen, P.H., Plovnick, R., Arno, A., Ivers, N., Quinn, G., Au, A., Johnston, R., Rada, G., Bagg, M., Jones, A., Ravaud, P., Boden, C., Kahale, L., Richter, B., Boisvert, I., Keshavarz, H., Ryan, R., Brandt, L., Kolakowsky-Hayner, S.A., Salama, D., Brazinova, A., Nagraj, S.K., Salanti, G., Buchbinder, R., Lasserson, T., Santaguida, L., Champion, C., Lawrence, R., Santesso, N., Chandler, J., Les, Z., Schünemann, H.J., Charidimou, A., Leucht, S., Shemilt, I., Chou, R., Low, N., Sherifali, D., Churchill, R., Maas, A., Siemieniuk, R., Cnossen, M.C., MacLehose, H., Simmonds, M., Cossi, M.-J., Macleod, M., Skoetz, N., Counotte, M., Marshall, I., Soares-Weiser, K., Craigie, S., Marshall, R., Srikanth, V., Dahm, P., Martin, N., Sullivan, K., Danilkewich, A., Martínez García, L., Synnot, A., Danko, K., Mavergames, C., Taylor, M., Donoghue, E., Maxwell, L.J., Thayer, K., Dressler, C., McAuley, J., Thomas, J., Egan, C., McDonald, Steve, Tritton, R., Elliott, J., McKenzie, J., Tsafnat, G., Elliott, S.A., Meerpohl, J., Tugwell, P., Etxeandia, I., Merner, B., Turgeon, A., Featherstone, R., Mondello, S., Turner, T., Foxlee, R., Morley, R., van Valkenhoef, G., Garner, P., Munafo, M., Vandvik, P., Gerrity, M., Munn, Z., Wallace, B., Glasziou, P., Murano, M., Wallace, S.A., Green, S., Newman, K., Watts,

C., Grimshaw, J., Nieuwlaat, R., Weeks, L., Gurusamy, K., Nikolakopoulou, A., Weigl, A., Haddaway, N., Noel-Storr, A., Wells, G., Hartling, L., O'Connor, A., Wiercioch, W., Hayden, J., Page, M., Wolfenden, L., Helfand, M., Pahwa, M., Yepes Nuñez, J.J., Higgins, J., Pardo, J.P., Yost, J., Hill, S., Pearson, L., 2017. Living systematic reviews: 2. Combining human and machine effort. Journal of Clinical Epidemiology 91, 31–37. https://doi.org/10.1016/j.jclinepi.2017.08.011

Tsafnat, G., Dunn, A., Glasziou, P., Coiera, E., 2013. The automation of systematic reviews. British Medical Journal 346, f139. https://doi.org/10.1136/bmj.f139

Tugwell, P., Haynes, R.B., 2006. Assessing claims of causation, in: Tugwell, B., Haynes, R.B., Sackett, D.L., Guyatt, G.H., Tugwell, P. (Eds.), Clinical epidemiology: how to do clinical practice research. The University of Chicago Press Philadelphia, Pennsylvania, pp. 356–387.

Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36, 1–48. https://doi.org/10.18637/jss.v036.i03

Wallace, B.C., Dahabreh, I.J., Schmid, C.H., Lau, J., Trikalinos, T.A., 2014. Modernizing Evidence Synthesis for Evidence-Based Medicine, in: Greenes, R.A. (Second Edition. (Ed.), Clinical Decision Support. Elsevier, Oxford, pp. 339–361. https://doi.org/10.1016/B978-0-12-398476-0.00012-9

Wauchope, H.S., 2020. Working with large-scale population trend data in ecology and conservation: methods and applications. University of Cambridge. https://doi.org/10.17863/CAM.59354

Wauchope, H.S., Amano, T., Sutherland, W.J., Johnston, A., 2019. When can we trust population trends? A method for quantifying the effects of sampling interval and duration. Methods in Ecology and Evolution 10, 2067–2078. https://doi.org/10.1111/2041-210X.13302

Webb, J.A., Nichols, S.J., Norris, R.H., Stewardson, M.J., Wealands, S.R., Lea, P., 2012. Ecological Responses to Flow Alteration: Assessing Causal Relationships with Eco Evidence. Wetlands 32, 203–213. https://doi.org/10.1007/s13157-011-0249-5

Westgate, A.J., Read, A.J., Cox, T.M., Schofield, T.D., Whitaker, B.R., Anderson, K.E., 1998. Monitoring a rehabilitated harbor porpoise using satellite telemetry. Marine Mammal Science 14, 599–604. https://doi.org/10.1111/j.1748-7692.1998.tb00746.x

Wolfe, D.A., Champ, M.A., Flemer, D.A., Mearns, A.J., 1987. Long-term biological data sets: Their role in research, monitoring, and management of estuarine and coastal marine systems. Estuaries 10, 181. https://doi.org/10.2307/1351847

Supplementary Information

Figure S1

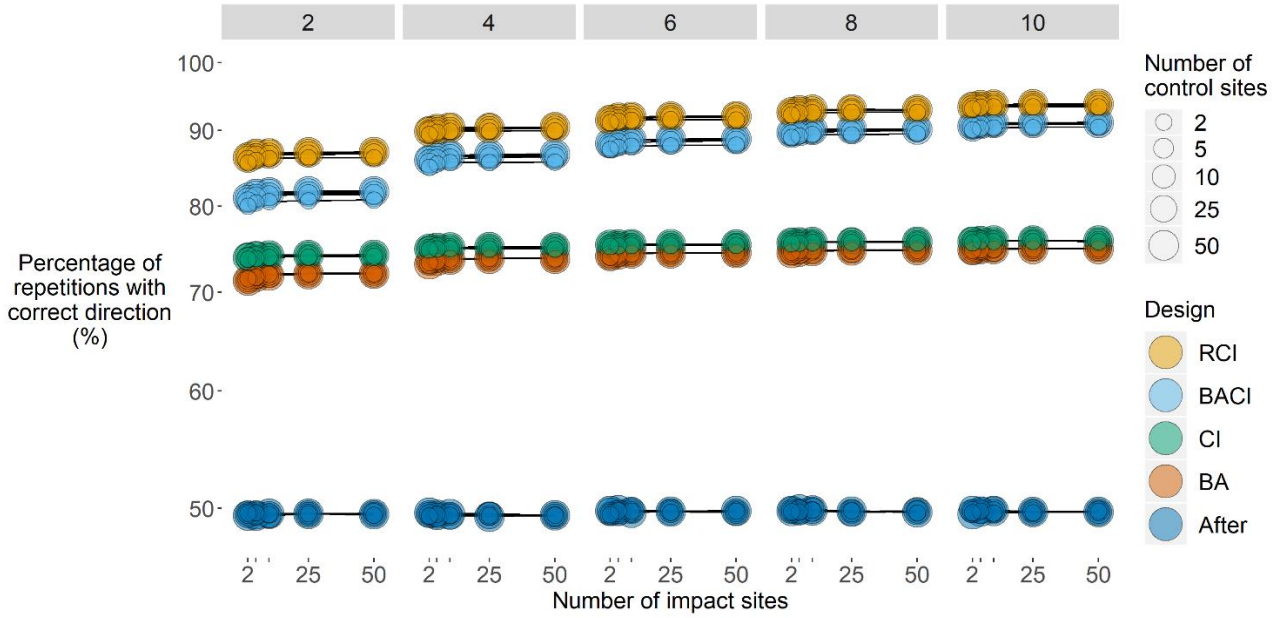


Figure S1 *– Percentage of simulation repetitions in which the design's effect size estimate had the correct direction. This is shown for multiple numbers of time steps simulated (T = 2, 4, 6, 8, or 10) and levels of spatial replication (control and impact sites separately). Circle size denotes the number of control sites. See Table 1 in main text for the definition of each design.*

Figure S2

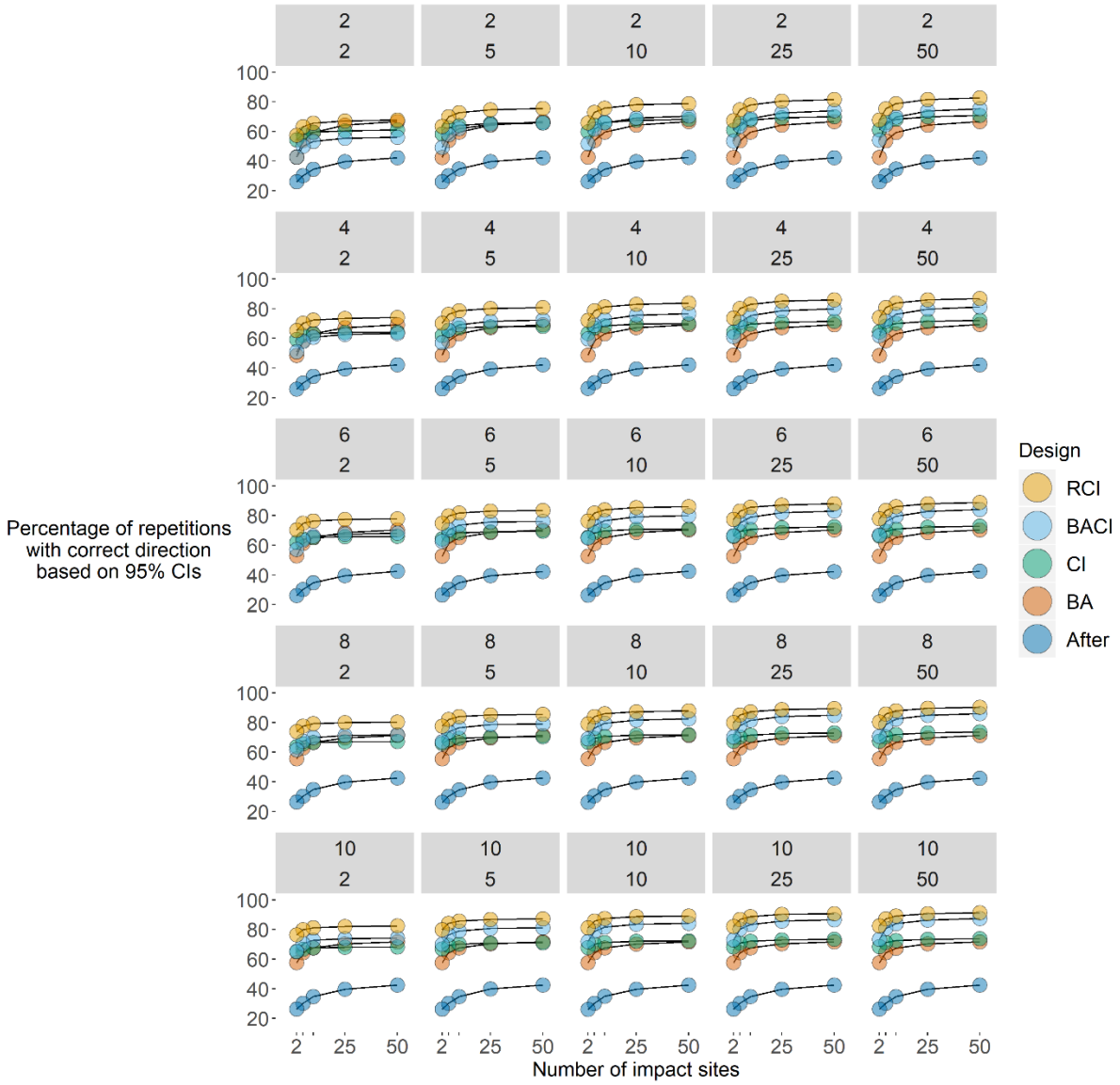


Figure S2 *– Percentage of simulation repetitions in which the entire range of uncertainty (95% Confidence Intervals) for the effect size estimate fell in the correct direction. This is shown for multiple numbers of time steps simulated (upper facet label – T = 2, 4, 6, 8, or 10) and levels of spatial replication (x-axis: impact sites – 2, 5, 10, 25, 50; lower facet label: control sites – 2, 5, 10, 25, 50). See Table 1 in main text for the definition of each design.*

Figure S3

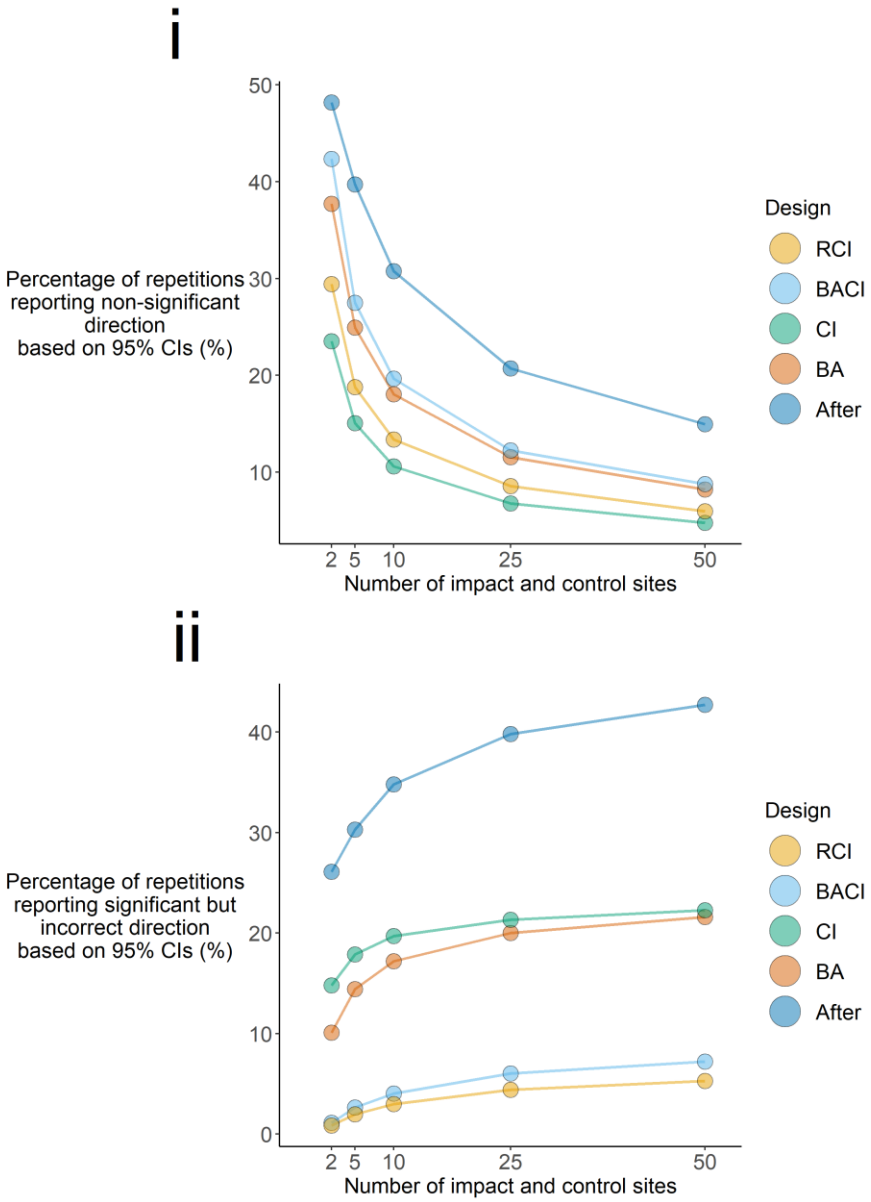


Figure S3 *– Percentage of simulation repetitions in which the design's the range of uncertainty (95% Confidence Intervals) for the effect size estimate overlapped with zero (Fig.S3-i; non-significant direction) or fell entirely in the wrong direction (Fig.S3-ii). This is shown for 6 time steps and equal numbers of control and impact sites. See Table 1 in main text for the definition of each design.*

Figure S4

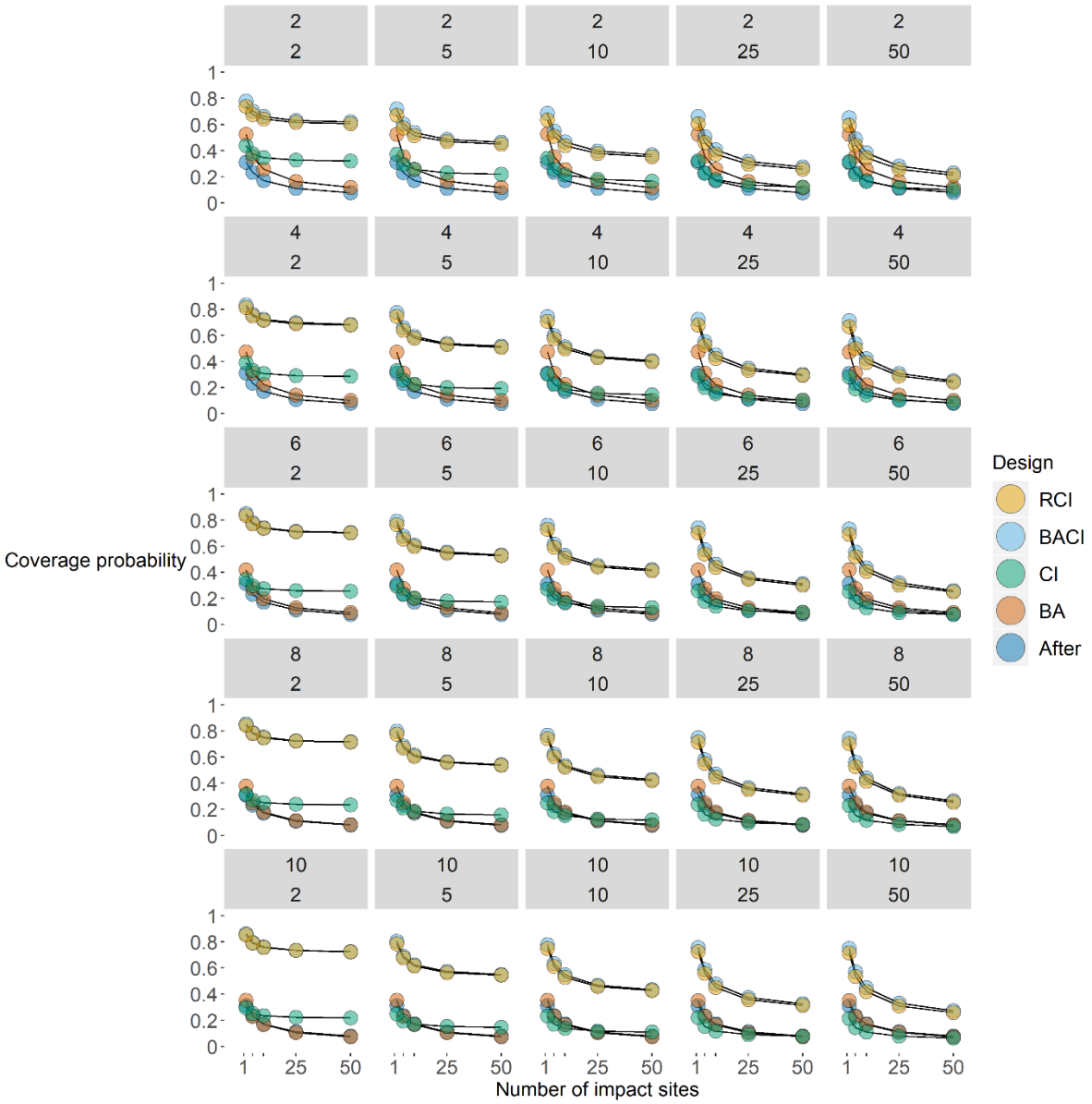


Figure S4 *– Coverage probabilities of designs (probability true effect lies within 95% Confidence Intervals of each design's effect size estimate). This is shown for multiple levels of spatial replication (x-axis: impact sites – 2, 5, 10, 25, 50; lower facet label: control sites – 2, 5, 10, 25, 50) and for multiple numbers of time steps simulated (upper facet label: T = 2, 4, 6, 8, or 10). See Table 1 in main text for the definition of each design.*

Figure S5

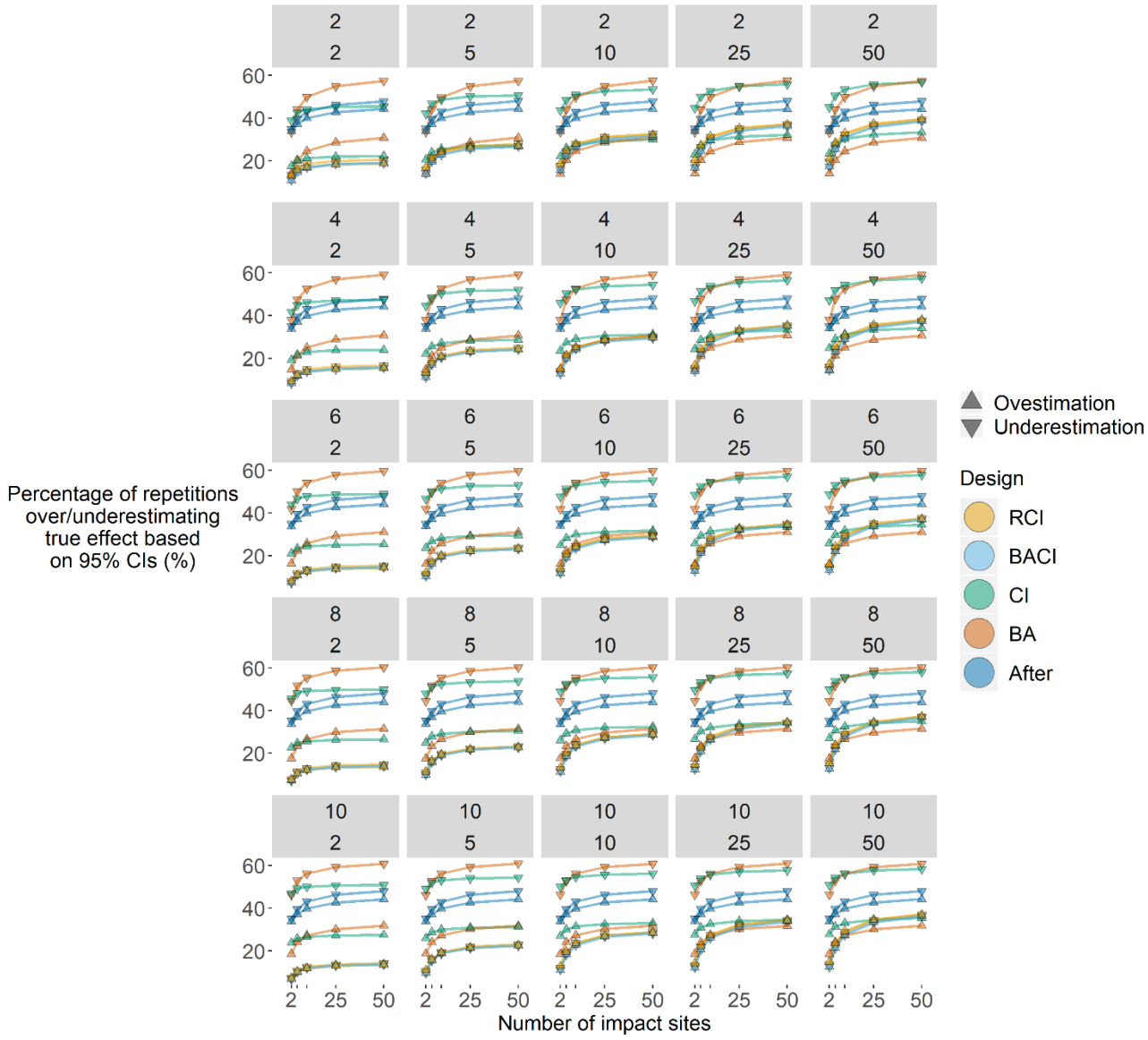


Figure S5 *– Percentage of simulation repetitions in which the entire range of uncertainty (95% Confidence Intervals) was either greater than (overestimate – upward triangle) or less than (underestimate – downward triangle) the true effect. This is shown for multiple levels of spatial replication (x-axis: impact sites – 2, 5, 10, 25, 50; lower facet label: control sites – 2, 5, 10, 25, 50) and for multiple numbers of time steps simulated (upper facet label: T = 2, 4, 6, 8, or 10). See Table 1 in main text for the definition of each design.*

Figure S6

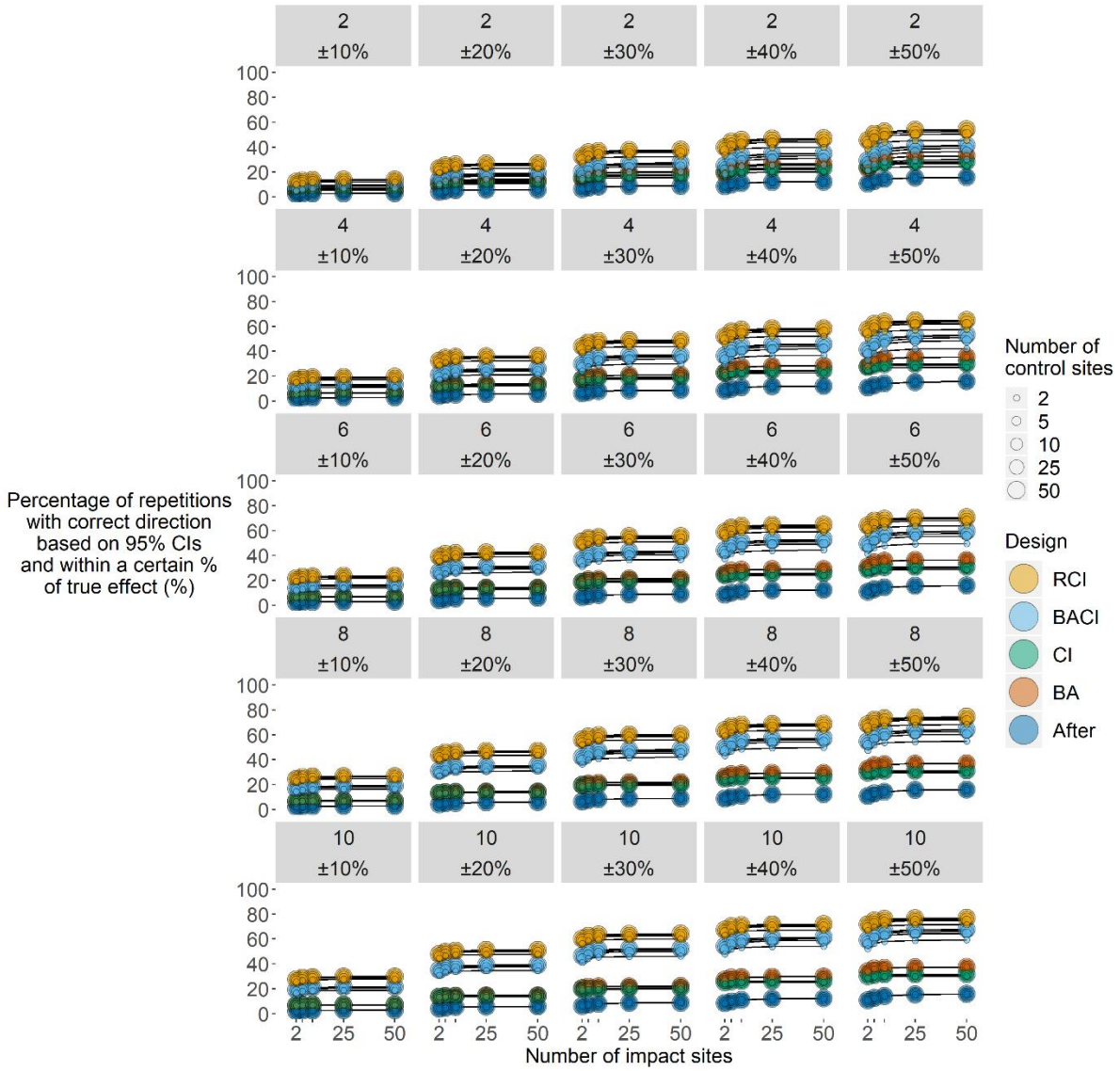


Figure S6 *– Percentage of simulation repetitions in which the design's effect size was both within 10, 30 or 50% of the true effect size and had the correct direction for multiple levels of spatial replication (control and impact sites separately). Graphs show all possible combinations of accuracy thresholds (±10%, ±30%, or ±50%) and time steps simulated (T = 2, 4, 6, 8, or 10). Circle size denotes the number of control sites. See Table 1 in main text for the definition of each design.*

Figure S7

Greater proportional change in control sites from before to after the impact (*C*) reduced the performance of BA designs substantially (Fig.S7). When there was no change in control sites, and consequently zero bias, BA designs performed as well as RCI designs. BACI, RCI and CI designs actually increased in performance as the overall size of the true effect increased with increasing proportional change in *C*; there was negligible change in the performance of After designs (Fig.S7).

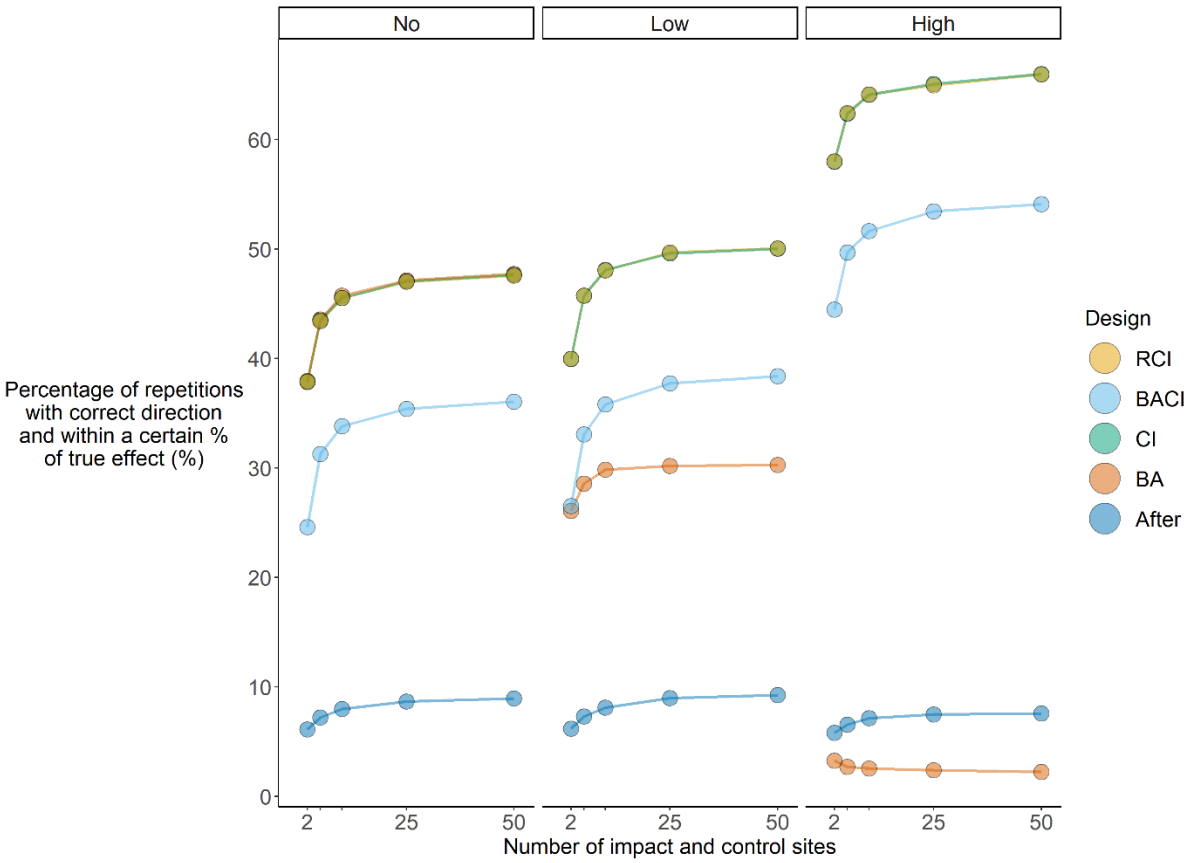


Figure S7 *– Percentage of simulation repetitions in which the design's effect size was both within 30% of the true effect size and had the correct direction for multiple levels of the proportional change in control sites from before to after the impact (No bias: C = 1, Low bias: C = 1.1 or 0.9, High bias: C = 1.3 or 0.7). This is presented for multiple numbers of control and impact sites. The number of time steps simulated was set at 6 and the proportional initial difference between impact and control sites (d_{CIB}) at 1 (i.e., no difference).*

Figure S8

Greater levels of the initial mean differences between control and impact groups in the before period (d_{CIB}) reduced the performance of CI designs substantially (Fig.S8). When there was no initial difference, CI designs performed as well as RCI designs; there was negligible change in the performance of any other design as the overall size of the true effect did not change (since this is not related to d_{CIB}; Fig.S8).

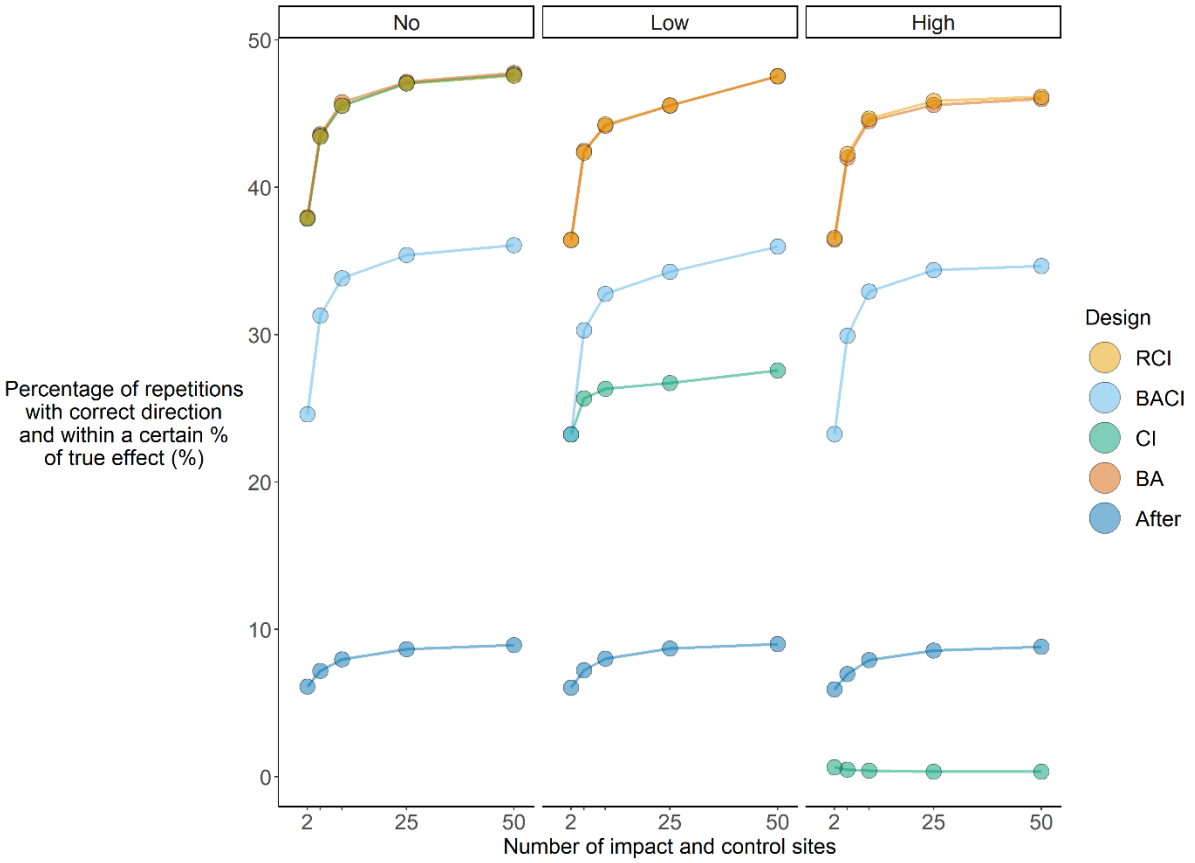


Figure S8 *– Percentage of simulation repetitions in which the design's effect size was both within 30% of the true effect size and had the correct direction for multiple levels of the proportional initial difference between impact and control sites (No bias: d_{CIB} = 1, Low bias: d_{CIB} = 1.1 or 0.9, High bias: d_{CIB} = 1.3 or 0.7). This is presented for multiple numbers of control and impact sites. The number of time steps simulated was set at 6 and proportional change in control sites from before to after the impact (C) at 1 (i.e., no change).*

Appendix S1

The major parameters in this simulation include: the change between the before and after periods in both impact sites (I) and control sites (C) and the difference between impact and control group means before intervention (d_{CIB}). The values of these three parameters were taken from 47 BACI datasets whereby the proportional change or difference (Table 3 in main text) was extracted for 2,002 effect sizes. We collated 45 datasets using a full search of the Web of Science with the search terms: ['BACI'] OR ['Before-After Control-Impact'] on the 18th December 2017. The search returned 674 results and we then refined this by selecting only 'Article' as a document type and only the following Web of Science Categories: 'Ecology', 'Marine Freshwater Biology', 'Biodiversity Conservation', 'Fisheries', 'Oceanography', 'Forestry', 'Zoology', Ornithology', 'Biology', 'Plant Sciences', 'Entomology', 'Remote Sensing', 'Toxicology' and 'Soil Science'. We were then left with 579 results. To be realistic about obtaining the data, we restricted the year of publication to 2002 (15 years prior to search), which reduced the number to 542. We then read the abstracts of all papers that we could access and excluded any studies that did not test the effect of an ecological intervention or threat using a BACI design with abundance, cover or density metrics of any taxon. This left 96 studies for which we then contacted the corresponding authors to ask for their raw data and received 47 datasets (25 quantified the impacts of threats and 22 quantified interventions).

These datasets may be affected by publication bias (i.e., that studies presenting statistically significant positive results are more likely to be published; Easterbrook et al. 1991). In this case, a BACI dataset is more likely to give a significant result if there are large, significant opposing changes in impact and control sites before versus after the impact (a time-impact interaction). Therefore, we might expect more BACI datasets to be published that show large, possibly unrepresentative changes in impacts and/or controls. To limit this bias, we extracted all the raw data from datasets, rather than just the published, focal data. We also tried to counteract extreme values from datasets by only considering data within the Interquartile Range (IQR) of each parameter, separately. These 25% and 75% quantiles became the minimum and maximum values of the subsetted data (number of values per parameter: I=1,320; C=1,206; d_{CIB}=1,166). We randomly sampled this subsetted data for each parameter 1,000 times with replacement (Table 3 in Results) to create 1,000 unique simulation scenarios.

References

Damgaard, C., 2019. A Critique of the Space-for-Time Substitution Practice in Community Ecology. Trends in Ecology and Evolution 34, 416–421.

Easterbrook, P.J., Berlin, J.A., Gopalan, R., Matthews, D.R., 1991. Publication bias in clinical research. Lancet 337, 86772.

Appendix S2

Calculating weights

Our accuracy weights take the following forms for each design using an accuracy threshold of ±30%:

$$\text{BACI} = \frac{1}{1+e^{-\left(-0.813+0.0792\cdot\ln\left(n_{Impact\ sites}\right)+0.0797\cdot\ln\left(n_{Control\ sites}\right)\right)}}$$

$$\text{RCI} = \frac{1}{1+e^{-\left(-0.144+0.0544\cdot\ln\left(n_{Impact\ sites}\right)+0.0533\cdot\ln\left(n_{Control\ sites}\right)\right)}}$$

$$\text{BA} = 0.208$$

$$\text{CI} = 0.193$$

$$\text{After} = 0.0800$$

Here are two examples:

1. França et al. (2016) used a BACI design, 29 impact and five control units and thus receives an accuracy weight of:

$$\frac{1}{1+e^{-(-0.813+0.0792\cdot\ln(29)+0.0797\cdot\ln(5))}} = 0.397$$

2. Potts et al. (2009) uses an RCI design with 12 impact and control sites and thus receives an accuracy weight of:

$$\frac{1}{1+e^{-(0.0423+0.0231\cdot\ln(12)+0.0236\cdot\ln(12))}} = 0.536$$

Use in meta-analysis

To apply these weights to a random-effects model (REM) we can use the metafor package in R (Viechtbauer 2010). Our approach was to modify the inverse-variance weights matrix using our weights using the following R code:

base_model <- rma.mv(yi, vi, random = 1|rand_eff, data = our_data)

M <- base_model$M ###extract the marginal variance-covariance matrix

W <- solve(M) ### Take the inverse of M to give the weights matrix

C <- diag(our_data$acc_wei) ### create a diagonal matrix of accuracy weights

WC <- sqrt(C)%% W %*% sqrt(C)* ### multiply weight matrix by accuracy weight matrix

aw_model <- rma.mv(yi, vi, W = WC, random = 1|rand_eff, data = our_data) ### run REM

where *yi* = effect size estimate of a study (e.g., SMD), *vi* = variance of effect size estimate, *WC* is the modified weight matrix using our accuracy weights, and *acc_wei* is a column within the dataframe *our_data*. There are similar arguments to *W* in other meta-analytical models in

R (and other statistical software) that take a vector of user-defined weights such as the weight matrix we modified with our accuracy weights. This methodology is preliminary and is likely to need refining and testing in future to ensure it is robust and unbiased. The results of 3 meta-analyses (Sandström et al. 2019, Bernes et al. 2018, Eales et al. 2018) which we applied our accuracy weights to can be found on Zenodo: https://doi.org/10.5281/zenodo.4437010. We show the number of studies with different designs for each summary effect size (Table S1).

Table S1 *– All details of numbers of studies of different designs for all 128 extracted summary effect sizes from the 3 meta-analyses. Rows highlighted in grey show the excluded summary effect size comparisons due to only one type of study design being present. Summary effect ID refers to summary effect results detailed in data archived on Zenodo.*

		Number of studies			
Meta-analysis name	**Summary effect ID**	**BACI**	**BA**	**CI**	**Total**
Sandström et al. 2019	S1	10	0	23	33
Sandström et al. 2019	S2	8	0	11	19
Sandström et al. 2019	S3	2	0	8	10
Sandström et al. 2019	S4	0	0	4	4
Sandström et al. 2019	S5	1	0	13	14
Sandström et al. 2019	S6	3	0	6	9
Sandström et al. 2019	S7	1	1	1	3
Sandström et al. 2019	S8	0	0	6	6
Sandström et al. 2019	S9	1	1	3	5
Sandström et al. 2019	S10	1	1	0	2
Sandström et al. 2019	S11	0	1	6	7
Sandström et al. 2019	S12	12	0	18	30
Sandström et al. 2019	S13	10	0	10	20
Sandström et al. 2019	S14	2	0	6	8
Sandström et al. 2019	S15	0	0	2	2
Sandström et al. 2019	S16	9	0	8	17
Sandström et al. 2019	S17	0	0	6	6
Sandström et al. 2019	S18	1	1	1	3
Sandström et al. 2019	S19	1	1	0	2
Bernes et al. 2018	B1	1	18	46	65
Bernes et al. 2018	B2	3	21	49	73
Bernes et al. 2018	B3	0	0	4	4
Bernes et al. 2018	B4	1	15	13	29
Bernes et al. 2018	B5	0	0	5	5
Bernes et al. 2018	B6	0	0	9	9
Bernes et al. 2018	B7	2	15	21	38
Bernes et al. 2018	B8	2	19	33	54
Bernes et al. 2018	B9	2	18	22	42
Bernes et al. 2018	B10	0	0	21	21
Bernes et al. 2018	B11	4	14	30	48

Meta-analysis name	Summary effect ID	BACI	BA	CI	Total
Bernes et al. 2018	B12	1	14	16	31
Bernes et al. 2018	B13	14	0	10	24
Bernes et al. 2018	B14	14	0	5	19
Bernes et al. 2018	B15	14	0	2	16
Bernes et al. 2018	B16	14	0	7	21
Bernes et al. 2018	B17	0	0	9	9
Bernes et al. 2018	B18	15	10	0	25
Bernes et al. 2018	B19	2	42	0	44
Bernes et al. 2018	B20	1	1	1	3
Bernes et al. 2018	B21	0	17	0	17
Bernes et al. 2018	B22	0	16	3	19
Bernes et al. 2018	B23	14	12	0	26
Bernes et al. 2018	B24	3	16	0	19
Bernes et al. 2018	B25	0	32	0	32
Bernes et al. 2018	B26	3	15	0	18
Bernes et al. 2018	B27	0	10	0	10
Bernes et al. 2018	B28	0	4	0	4
Bernes et al. 2018	B29	0	8	1	9
Bernes et al. 2018	B30	17	0	0	17
Bernes et al. 2018	B31	10	0	0	10
Bernes et al. 2018	B32	2	0	0	2
Bernes et al. 2018	B33	2	0	0	2
Bernes et al. 2018	B34	5	0	0	5
Bernes et al. 2018	B35	8	0	0	8
Bernes et al. 2018	B36	11	2	0	13
Bernes et al. 2018	B37	6	0	0	6
Bernes et al. 2018	B38	5	1	0	6
Bernes et al. 2018	B39	3	0	1	4
Bernes et al. 2018	B40	3	0	0	3
Bernes et al. 2018	B41	4	0	1	5
Bernes et al. 2018	B42	18	4	0	22
Bernes et al. 2018	B43	10	4	0	14
Bernes et al. 2018	B44	26	10	0	36
Bernes et al. 2018	B45	4	0	0	4
Bernes et al. 2018	B46	14	6	1	21
Bernes et al. 2018	B47	2	2	0	4
Bernes et al. 2018	B48	33	12	0	45
Bernes et al. 2018	B49	5	1	3	9
Bernes et al. 2018	B50	10	6	0	16
Bernes et al. 2018	B51	1	2	0	3
Bernes et al. 2018	B52	25	11	0	36
Bernes et al. 2018	B53	1	0	2	3
Bernes et al. 2018	B54	7	6	0	13
Bernes et al. 2018	B55	16	10	0	26

Meta-analysis name	Summary effect ID	BACI	BA	CI	Total
Bernes et al. 2018	B56	0	0	2	2
Bernes et al. 2018	B57	6	6	0	12
Bernes et al. 2018	B58	21	2	0	23
Bernes et al. 2018	B59	10	8	0	18
Bernes et al. 2018	B60	14	6	1	21
Bernes et al. 2018	B61	1	2	0	3
Bernes et al. 2018	B62	8	0	1	9
Bernes et al. 2018	B63	38	18	0	56
Bernes et al. 2018	B64	1	1	0	2
Bernes et al. 2018	B65	5	14	0	19
Bernes et al. 2018	B66	6	1	0	7
Bernes et al. 2018	B67	24	0	0	24
Bernes et al. 2018	B68	4	2	0	6
Bernes et al. 2018	B69	6	0	1	7
Bernes et al. 2018	B70	14	14	0	28
Bernes et al. 2018	B71	64	17	3	84
Bernes et al. 2018	B72	25	7	1	33
Bernes et al. 2018	B73	9	14	0	23
Bernes et al. 2018	B74	57	17	3	77
Bernes et al. 2018	B75	37	7	1	45
Bernes et al. 2018	B76	15	0	0	15
Bernes et al. 2018	B77	18	0	0	18
Bernes et al. 2018	B78	9	0	0	9
Bernes et al. 2018	B79	13	6	0	19
Bernes et al. 2018	B80	2	0	2	4
Bernes et al. 2018	B81	15	6	2	23
Bernes et al. 2018	B82	13	6	0	19
Bernes et al. 2018	B83	1	0	1	2
Bernes et al. 2018	B84	2	6	0	8
Bernes et al. 2018	B85	2	6	0	8
Bernes et al. 2018	B86	5	6	0	11
Bernes et al. 2018	B87	2	6	0	8
Bernes et al. 2018	B88	13	0	2	15
Bernes et al. 2018	B89	2	6	0	8
Bernes et al. 2018	B90	15	6	2	23
Bernes et al. 2018	B91	10	0	1	11
Bernes et al. 2018	B92	20	14	2	36
Bernes et al. 2018	B93	5	0	0	5
Bernes et al. 2018	B94	4	14	1	19
Bernes et al. 2018	B95	6	0	0	6
Bernes et al. 2018	B96	10	0	2	12
Bernes et al. 2018	B97	4	0	1	5
Bernes et al. 2018	B98	5	0	0	5
Bernes et al. 2018	B99	6	0	0	6

Meta-analysis name	Summary effect ID	BACI	BA	CI	Total
Bernes et al. 2018	B100	6	0	0	6
Eales et al. 2018	E1	27	5	31	63
Eales et al. 2018	E2	5	0	5	10
Eales et al. 2018	E3	7	3	12	22
Eales et al. 2018	E4	10	2	11	23
Eales et al. 2018	E5	5	2	6	13
Eales et al. 2018	E6	8	3	2	13
Eales et al. 2018	E7	3	1	1	5
Eales et al. 2018	E8	4	0	2	6
Eales et al. 2018	E9	4	0	6	10

Other possible uses in evidence assessment

After assigning each study a weight, the sum, mean or median weight for a set of studies could be used to indicate the relative strength of the evidence as a whole. For example, we could set categories for strong, moderate, and weak evidence for studies: strong evidence = mean/median weight ≥0.4, sum weight > =2; moderate evidence = mean/median weight <0.4 and ≥0.2, sum weight >=1; weak evidence <0.2, sum weight <1. Systematic reviews or meta-analyses could also be weighted as single pieces of evidence themselves using this method, either by totalling or by finding the median or mean weight of studies they include. There are many possible ways this could be done, and further work is needed to justify and find appropriate ways to covert these scores from a continuous scale to a categorical one for use in evidence assessment.

References

França, F., Louzada, J., Korasaki, V., Griffiths, H., Silveira, J.M., Barlow, J., 2016. Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. Journal of Applied Ecology 53, 1098–1105.

Potts, S.G., Woodcock, B.A., Roberts, S.P.M., Tscheulin, T., Pilgrim, E.S., Brown, V.K., Tallowin, J.R., 2009. Enhancing pollinator biodiversity in intensive grasslands. Journal of Applied Ecology 46, 369–379.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36, 1–48.

Eales, J., Haddaway, N.R., Bernes, C., Cooke, S.J., Jonsson, B.G., Kouki, J., Petrokofsky, G., Taylor, J.J., 2018. What is the effect of prescribed burning in temperate and boreal forest on biodiversity, beyond pyrophilous and saproxylic species? A systematic review. Environmental Evidence 7, 19.

Bernes, C., Macura, B., Jonsson, B.G., Junninen, K., Müller, J., Sandström, J., Lõhmus, A., Macdonald, E., 2018. Manipulating ungulate herbivory in temperate and boreal forests: effects on vegetation and invertebrates. A systematic review. Environmental Evidence 7, 13.

Sandström, J., Bernes, C., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., Jonsson, B.G., 2019. Impacts of dead wood manipulation on the biodiversity of temperate and boreal forests. A systematic review. Journal of Applied Ecology 56, 1770–1781.

Appendix S3

Here we present the equivalent results, accuracy weights and equations (Tables S1 and S4) as those in the main text, but instead for the percentage of simulation repetitions where the estimated effect size gave the correct direction and was within ±10% or ±50% of the true effect. We also present associated tables (Tables S2-3 and S5-6) to show how we selected the best beta GLM model for generating these equations and weights.

±10% threshold

BACI and RCI models with both impact sites and control sites as predictor variables were selected as the best models because they had the lowest values of AIC; although the model using an interaction term between impact and control sites was within 2 units of AIC, we selected the models without an interaction term because they were more parsimonious (Tables S1-S3). Both selected models were only slightly over-dispersed (RCI model: θ = 1.19; BACI model: θ = 1.19) and Pearson's χ2 residuals were non-significant (p>0.05) suggesting no significant patterns remained in the residuals. There were also no observable patterns between residuals and explanatory variables or fitted values.

Table S1 *– Results of Generalised Linear Models for BACI and RCI designs and accuracy weight equations for all designs based on data for the ±10% accuracy threshold. Coefficients are in log odds (3.s.f.).* n_I *= number of independent impact units,* n_C *= number of independent control units.*

	Intercept		**Number of impact sites** $\ln(n_I)$		**Number of control sites** $\ln(n_C)$		**Quasi- R^2**
Design	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**
BACI	−2.01	0.0256	0.0556	0.00714	0.0580	0.00714	0.838
RCI	−1.42	0.0145	0.0370	0.00409	0.0359	0.00409	0.865
	Accuracy weights/equations						
BACI	$\frac{1}{1+e^{-(-2.01+0.0556\cdot\ln(n_{Impact\ sites})+0.0580\cdot\ln(n_{Control\ sites}))}}$						
RCI	$\frac{1}{1+e^{-(-1.42+0.0370\cdot\ln(n_{Impact\ sites})+0.0359\cdot\ln(n_{Control\ sites}))}}$						
BA	0.0686						
CI	0.0681						
After	0.0261						

Table S2 *– Models considered in model selection process for finding weighting equation for RCI design.* N_I *= Number of Impact sites,* N_C *= Number of Control sites.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-196.7415
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-194.7432
C	$ln(N_C)$	-162.3059
D	$ln(N_I)$	-163.4270

Table S3 *– Models considered in model selection process for finding weighting equation for BACI design.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-184.4541
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-182.4724
C	$ln(N_C)$	-155.3550
D	$ln(N_I)$	-153.8156

±50% threshold

BACI and RCI models with both impact sites and control sites as predictor variables were selected as the best models because they had the lowest values of AIC; although the model using an interaction term between impact and control sites was within 2 units of AIC, we selected the models without an interaction term because they were more parsimonious (Tables S4-6). Both selected models were only slightly over-dispersed (RCI model: θ = 1.19; BACI model: θ = 1.19) and Pearson's χ2 residuals were non-significant (p>0.05) suggesting no significant patterns remained in the residuals. There were also no observable patterns between residuals and explanatory variables or fitted values.

Table S4 *– Results of Generalised Linear Models for BACI and RCI designs and accuracy weight equations for all designs based on data from Fig.4 in main text and Fig.S3. Coefficients are in log odds (3.s.f.).* n_I *= number of independent impact units,* n_C *= number of independent control units.*

	Intercept		**Number of impact sites** $\ln(\boldsymbol{n_I})$		**Number of control sites** $\ln(\boldsymbol{n_C})$		**Quasi- R^2**
Design	Coef.	SE	Coef.	SE	Coef.	SE	Coef.
BACI	-0.404	0.0464	0.116	0.0132	0.112	0.0132	0.971
RCI	0.286	0.0345	0.0844	0.00994	0.0854	0.00994	0.855
	Accuracy weights/equations						
BACI	$\frac{1}{1+e^{-(-0.404+0.116\cdot\ln(n_{Impact\ sites})+0.112\cdot\ln(n_{Control\ sites}))}}$						
RCI	$\frac{1}{1+e^{-(0.286+0.0844\cdot\ln(n_{Impact\ sites})+0.0854\cdot\ln(n_{Control\ sites}))}}$						
BA	0.348						
CI	0.296						
After	0.140						

Table S5 *– Models considered in model selection process for finding weighting equation for RCI design.* N_I *= Number of Impact sites,* N_C *= Number of Control sites.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-139.8733
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-138.3100
C	$ln(N_C)$	-107.6753
D	$ln(N_I)$	-108.1101

Table S6 *– Models considered in model selection process for finding weighting equation for BACI design.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-120.35132
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-118.97371
C	$ln(N_C)$	-87.26081
D	$ln(N_I)$	-88.44984

Appendix S4

Table S1 *– Results of Generalised Linear Models for BACI and RCI designs and accuracy weight equations for all designs based on data for the 30% accuracy threshold from Fig.4 in main text, Fig.S3, and Equations 2 and 3. Coefficients are in log odds (3.s.f.). n_I = number of independent impact units, n_C = number of independent control units. For equations and weights for other accuracy thresholds see Appendix S3.*

	Intercept		**Number of impact sites $\ln(n_I)$**		**Number of control sites $\ln(n_C)$**		**Quasi- R^2**
Design	Coef.	SE	Coef.	SE	Coef.	SE	Coef.
BACI	−0.813	0.0340	0.0792	0.00960	0.0797	0.00960	0.847
RCI	−0.144	0.0226	0.0544	0.0646	0.0533	0.0646	0.847
	Accuracy weights/equations						
BACI	$\frac{1}{1+e^{-(-0.813+0.0792\cdot\ln(n_{Impact\ sites})+0.0797\cdot\ln(n_{Control\ sites}))}}$						
RCI	$\frac{1}{1+e^{-(-0.144+0.0544\cdot\ln(n_{Impact\ sites})+0.0533\cdot\ln(n_{Control\ sites}))}}$						
BA	0.208						
CI	0.193						
After	0.0800						

Table S2 *– Models considered in model selection process for finding weighting equation for RCI design.* N_I *= Number of Impact sites,* N_C *= Number of Control sites.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-155.7721
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-153.7780
C	$ln(N_C)$	-124.1265
D	$ln(N_I)$	-124.8542

Table S3 *– Models considered in model selection process for finding weighting equation for BACI design.* N_I *= Number of Impact sites,* N_C *= Number of Control sites.*

Model	**Parameters**	**AIC**
A (Best model)	$ln(N_I) + ln(N_C)$	-138.3087
B	$ln(N_I) + ln(N_C) + ln(N_I) * ln(N_C)$	-136.3896
C	$ln(N_C)$	-107.2681
D	$ln(N_I)$	-107.0637

Appendix S5

Datasets that are openly accessible and available online that we used to parameterise simulations are listed below.

Bernes, C., Macura, B., Jonsson, B.G., Junninen, K., Müller, J., Sandström, J., Lõhmus, A., Macdonald, E., 2018. Manipulating ungulate herbivory in temperate and boreal forests: effects on vegetation and invertebrates. A systematic review. Environmental Evidence 7, 13. https://doi.org/10.1186/s13750-018-0125-3

Burge, O.R., Bodmin, K.A., Clarkson, B.R., Bartlam, S., Watts, C.H., Tanner, C.C., 2017. Glyphosate redirects wetland vegetation trajectory following willow invasion. Applied Vegetation Science 20, 620–630.

Dietl, G.P., Durham, S.R., 2016. Data from: Geohistorical records indicate no impact of the Deepwater Horizon oil spill on oyster body size. Dryad Digital Repository. https://doi.org/10.5061/dryad.bc80t

Eales, J., Haddaway, N.R., Bernes, C., Cooke, S.J., Jonsson, B.G., Kouki, J., Petrokofsky, G., Taylor, J.J., 2018. What is the effect of prescribed burning in temperate and boreal forest on biodiversity, beyond pyrophilous and saproxylic species? A systematic review. Environmental Evidence 7, 19. https://doi.org/10.1186/s13750-018-0131-5

Moland, E., Olsen, E.M., Knutsen, H., Garrigou, P., Espeland, S.H., Kleiven, A.R., André, C., Knutsen, J.A., 2013. Lobster and cod benefit from small-scale northern marine protected areas: inference from an empirical before–after control-impact study. Proceedings of the Royal Society B: Biological Sciences 280, 20122679.

Sandström, J., Bernes, C., Junninen, K., Lõhmus, A., Macdonald, E., Müller, J., Jonsson, B.G., 2019. Impacts of dead wood manipulation on the biodiversity of temperate and boreal forests. A systematic review. Journal of Applied Ecology 56, 1770–1781. https://doi.org/10.1111/1365-2664.13395

Sepúlveda, R.D., Valdivia, N., 2016. Localised effects of a mega-disturbance: spatiotemporal responses of intertidal sandy shore communities to the 2010 Chilean earthquake. PLOS ONE 11, e0157910.

Williams, D.E., Miller, M.W., Bright, A.J., Cameron, C.M., 2014. Removal of corallivorous snails as a proactive tool for the conservation of acroporid corals. PeerJ 2, e680. https://doi.org/10.7717/peerj.680

3 | Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences

This chapter was published as:

Christie, A.P., Abecasis, D., Adjeroud, M., Alonso, J.C., Amano, T., Anton, A., Baldigo, B.P., Barrientos, R., Bicknell, J.E., Buhl, D.A., Cebrian, J., Ceia, R.S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M.D., Davoult, D., De Backer, A., Donovan, M.K., Eddy, T.D., França, F.M., Gardner, J.P.A., Harris, B.P., Huusko, A., Jones, I.L., Kelaher, B.P., Kotiaho, J.S., López-Baucells, A., Major, H.L., Mäki-Petäys, A., Martín, B., Martín, C.A., Martin, P.A., Mateos-Molina, D., McConnaughey, R.A., Meroni, M., Meyer, C.F.J., Mills, K., Montefalcone, M., Noreika, N., Palacín, C., Pande, A., Pitcher, C.R., Ponce, C., Rinella, M., Rocha, R., Ruiz-Delgado, M.C., Schmitter-Soto, J.J., Shaffer, J.A., Sharma, S., Sher, A.A., Stagnol, D., Stanley, T.R., Stokesbury, K.D.E., Torres, A., Tully, O., Vehanen, T., Watts, C., Zhao, Q., Sutherland, W.J., 2020. Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. Nature Communications 11, 6377. https://doi.org/10.1038/s41467-020-20142-y

Abstract

Building trust in science and evidence-based decision-making depends heavily on the reliability of studies and their findings. Researchers employ many different study designs that vary in their risk of bias to evaluate the true effect of interventions or impacts. Here, we empirically quantify, on a large scale, the prevalence of different study designs and the magnitude of bias in their estimates. Randomised designs and controlled observational designs with pre-intervention sampling were used by just 23% of intervention studies in biodiversity conservation, and 36% of intervention studies in social science. We demonstrate, through pairwise within-study comparisons across 49 environmental datasets, that these types of designs usually give less biased estimates than simpler observational designs. We propose a model-based approach to combine study estimates that may suffer from different levels of study design bias, discuss the implications for evidence synthesis, and how to facilitate the use of more credible study designs.

Introduction

The ability of science to reliably guide evidence-based decision-making hinges on the accuracy and reliability of studies and their results (Donnelly et al., 2018; McKinnon et al., 2015). Well-designed, randomised experiments are widely accepted to yield more credible results than non-randomised, 'observational studies' that attempt to approximate and mimic randomised experiments (Rubin, 2008). Randomisation is a key element of study design that is widely used across many disciplines because of its ability to remove confounding biases (through random assignment of the treatment or impact of interest; Fisher, 1925; Peirce and Jastrow, 1884). However, ethical, logistical, and economic constraints often prevent the implementation of randomised experiments, whereas non-randomised observational studies have become popular as they take advantage of historical data for new research questions, larger sample sizes, less costly implementation, and more relevant and representative study systems or populations (Angrist and Pischke, 2008; de Palma et al., 2018; Sagarin and Pauchard, 2010; Shadish et al., 2002). Observational studies nevertheless face the challenge of accounting for confounding biases without randomisation, which has led to innovations in study design.

We define 'study design' as an organised way of collecting data. Importantly, we distinguish between data collection and statistical analysis (as opposed to other authors; Rosenbaum, 2010) because of the belief that bias introduced by a flawed design is often much more important than bias introduced by statistical analyses. This was emphasised by Light et al. (1990, p.5): "You can't fix by analysis what you bungled by design..."; and Rubin (2008): "Design trumps analysis." Nevertheless, the importance of study design has often been overlooked in debates over the inability of researchers to reproduce the original results of published studies (so-called 'reproducibility crises'; Ioannidis, 2005; Open Science Collaboration, 2015) as more attention is typically devoted to issues such as p-hacking (John et al., 2012) or Hypothesising After Results are Known ('HARKing'; Kerr, 1998).

To demonstrate the importance of study designs, we can use the following decomposition of estimation error equation (Zhao et al., 2019):

$$\text{Estimation error} = (\text{Estimator} - \text{True causal effect}) = (\text{Design bias} + \text{Modelling bias} + \text{Statistical noise}) \quad \text{(Equation 1)}.$$

This demonstrates that even if we improve the quality of modelling and analysis (to reduce modelling bias through a better bias-variance trade-off; Friedman et al., 2001) or increase sample size (to reduce statistical noise), we cannot remove the intrinsic bias introduced by the choice of study design (design bias) unless we collect the data in a different way. The

importance of study design in determining the levels of bias in study results therefore cannot be overstated.

For the purposes of this study, we consider six commonly used study designs; differences and connections can be visualised in Figure 1. There are three major components that allow us to define these designs: randomisation, sampling before and after the impact of interest occurs, and the use of a control group.

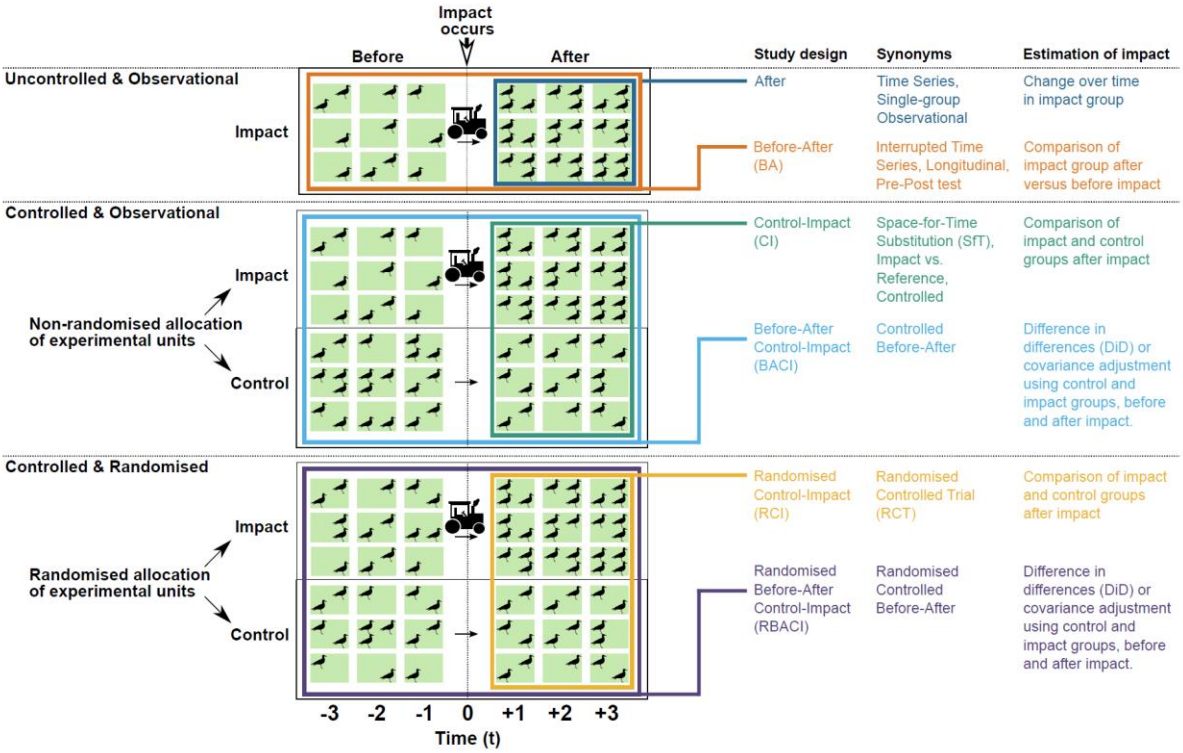


Figure 1 *– Comparison of different study designs used to evaluate the effect of an impact. A hypothetical study set-up is shown where the abundance of birds in three impact and control replicates (e.g., fields represented by blocks in a row) are monitored before and after an impact (e.g., ploughing) that occurs in year zero. Different colours represent each study design and illustrate how replicates are sampled. Approaches for calculating an estimate of the impact for each design are also shown, along with synonyms from different disciplines.*

Of the non-randomised observational designs, the Before-After Control-Impact (BACI) design uses a control group and samples before and after the impact occurs (i.e., in the 'before-period' and the 'after-period'). Its rationale is to explicitly account for pre-existing differences between the impact group (exposed to the impact) and control group in the before-period, which might otherwise bias the estimate of the impact's true effect (Angrist and Pischke, 2008; Stewart-Oaten and Bence, 2001; Underwood, 1991).

The BACI design improves upon several other commonly used observational study designs, of which there are two uncontrolled designs: After, and Before-After (BA). An After design

monitors an impact group in the after-period, whilst a BA design compares the state of the impact group between the before- and after-periods. Both designs can be expected to yield poor estimates of the impact's true effect (large design bias; Equation 1) because changes in the response variable could have occurred without the impact (e.g., due to natural seasonal changes; Fig.1).

The other observational design is Control-Impact (CI), which compares the impact group and control group in the after-period (Fig.1). This design may suffer from design bias introduced by pre-existing differences between the impact group and control group in the before-period; bias that the BACI design was developed to account for (Eddy et al., 2014; Sher et al., 2018). These differences have many possible sources, including experimenter bias, logistical and environmental constraints, and various confounding factors (variables that change the propensity of receiving the impact), but can be adjusted for using certain data pre-processing techniques such as matching and stratification (Imbens and Rubin, 2015).

Among the randomised designs, the most commonly used are counterparts to the observational CI and BACI designs: Randomised Control-Impact (R-CI) and Randomised Before-After Control-Impact (R-BACI) designs. The R-CI design, often termed 'Randomised Controlled Trials' (RCTs) in medicine and hailed as the 'gold standard' (Greenhalgh, 2019; Salmond, 2008), removes any pre-impact differences in a stochastic sense, resulting in zero design bias (Equation 1). Similarly, the R-BACI design should also have zero design bias, and the impact group measurements in the before-period could be used to improve the efficiency of the statistical estimator. No randomised equivalents exist of After or BA designs as they are uncontrolled.

It is important to briefly note that there is debate over two major statistical methods that can be used to analyse data collected using BACI and R-BACI designs, and which is superior at reducing modelling bias (Geijzendorffer et al., 2017; Equation 1). These statistical methods are: i.) Differences in Differences (DiD) estimator; and ii.) covariance adjustment using the before-period response, which is an extension of Analysis of Covariance (ANCOVA) for generalised linear models — herein termed 'covariance adjustment' (Fig.1). These estimators rely on different assumptions to obtain unbiased estimates of the impact's true effect. The DiD estimator assumes that the control group response accurately represents the impact group response had it not been exposed to the impact ('parallel trends'; Dimick and Ryan, 2014; Underwood, 1991), whereas covariance adjustment assumes there are no unmeasured confounders and linear model assumptions hold (Angrist and Pischke, 2008; Ding and Li, 2019).

With similar sample sizes, randomised designs (R-BACI and R-CI) are expected (based on Equation 1) to be less biased than controlled, observational designs with sampling in the before-period (BACI), which in turn should be superior to observational designs without sampling in the before-period (CI) or without a control group (BA and After designs (Christie et al., 2019; de Palma et al., 2018). Between randomised designs, we might expect that an R-BACI design performs better than a R-CI design because utilising extra data before the impact may improve the efficiency of the statistical estimator by explicitly characterising pre-existing differences between the impact group and control group.

Given the likely differences in bias associated with different study designs, concerns have been raised over the use of poorly designed studies in several scientific disciplines (Christie et al., 2020a, 2020b; de Palma et al., 2018; Goldenhar and Schulte, 1994; Junker et al., 2020; Kilkenny et al., 2009; Moscoe et al., 2015; Watson et al., 2019). Some disciplines, such as the social and medical sciences, commonly undertake direct comparisons of results obtained by randomised and non-randomised designs within a single study (Altindag et al., 2019; Chaplin et al., 2018; Cook et al., 2008) or between multiple studies (between-study comparisons; Benson and Hartz, 2000; dos Santos Ribas et al., 2020; Ioannidis et al., 2001) to specifically understand the influence of study designs on research findings. However, within-study comparisons are limited in their scope (e.g., a single study; França et al., 2016; Smokorowski et al., 2017) and between-study comparisons can be confounded by variability in context or study populations (Duvendack et al., 2012). Overall, we lack quantitative estimates of the prevalence of different study designs and the levels of bias associated with their results.

In this work, we aim to first quantify the prevalence of different study designs in the social and environmental sciences. To fill this knowledge gap, we take advantage of summaries for several thousand biodiversity conservation intervention studies in the Conservation Evidence database (Sutherland et al., 2019; www.conservationevidence.com) and social intervention studies in systematic reviews by the Campbell Collaboration (www.campbellcollaboration .org). We then quantify the levels of bias in estimates obtained by different study designs (R-BACI, R-CI, BACI, BA, and CI) by applying a hierarchical model to approximately 1,000 within-study comparisons across 49 raw environmental datasets from a range of fields. We show that R-BACI, R-CI, and BACI designs are poorly represented in studies testing biodiversity conservation and social interventions, and that these types of designs tend to give less biased estimates than simpler observational designs. We propose a model-based approach to combine study estimates that may suffer from different levels of design bias, discuss the implications for evidence synthesis, and how to facilitate the use of more reliable designs.

Materials and methods

Quantifying the use of different designs

We compared the use of different study designs in the literature that quantitatively tested interventions between the fields of biodiversity conservation (4,260 studies collated by Conservation Evidence; Sutherland et al., 2019) and social science (1,009 studies found by 32 systematic reviews produced by the Campbell Collaboration: www.campbellcollaboration.org).

Conservation Evidence is a database of intervention studies, each of which has quantitatively tested a conservation intervention (e.g., sowing strips of wildflower seeds on farmland to benefit birds), that is continuously being updated through comprehensive, manual searches of conservation journals for a wide range of fields in biodiversity conservation (e.g., amphibian, bird, peatland, and farmland conservation; Sutherland et al., 2019). To obtain the proportion of studies with each design from Conservation Evidence, we simply extracted the type of study design used by each study from the database in 2019 – the study design was determined using a standardised set of criteria; reviews were not included (Table 1). We checked if the designs reported in the database accurately reflected the designs in the original publication and found that for a random subset of 356 studies, 95.1% were accurately described.

Each systematic review produced by the Campbell Collaboration collates and analyses studies that test a specific social intervention; we collated reviews that tested a variety of social interventions across several fields in the social sciences, including education, crime and justice, international development, and social welfare (Appendix S1). We retrieved systematic reviews produced by the Campbell Collaboration by searching their website (www.campbellcollaboration.org) for reviews published between 2013–2019 (as of 8th September 2019) — we limited the date range as we could not go through every review. As we were interested in the use of study designs in the wider social-science literature, we only considered reviews (32 in total) that contained sufficient information on the number of included and excluded studies that used different study designs. Studies may be excluded from systematic reviews for several reasons, such as their relevance to the scope of the review (e.g., testing a relevant intervention) and their study design. We only considered studies if the sole reason for their exclusion from the review was their study design – i.e., reviews clearly reported that the study was excluded because it used a particular study design, and not because of any other reason, such as its relevance to the review's research questions. We calculated the proportion of studies that used each design in each systematic review (using the same criteria as for the biodiversity-conservation literature – see Table 1) and then averaged these proportions across all reviews.

Table 1 *– Definitions used to categorise studies based on the study design they used. See also Figure 1 for visual illustration and comparison of designs. Reviews from the Conservation Evidence database were not included.*

Study design	Controlled?	Sampling before impact occurs?	Randomised allocation of replicates to the impact group and control group?
After	No	No	No
Before-After (BA)	No	Yes	No
Control-Impact (CI)	Yes	No	No
Before-After Control-Impact (BACI)	Yes	Yes	No
Randomised Control-Impact (R-CI)	Yes	No	Yes
Randomised Before-After Control-Impact (R-BACI)	Yes	Yes	Yes

Within-study comparisons of different study designs

We wanted to make direct within-study comparisons between the estimates obtained by different study designs (e.g., see Cook et al., 2008, LaLonde, 1986, Long et al., 2008 for single within-study comparisons) for many different studies. If a dataset contains data collected using a BACI design, subsets of these data can be used to mimic the use of other study designs (a BA design using only data for the impact group, and a CI design using only data collected after the impact occurred). Similarly, if data were collected using a R-BACI design, subsets of these data can be used to mimic the use of a BA design and a R-CI design. Collecting BACI and R-BACI datasets would therefore allow us to make direct within-study comparisons of the estimates obtained by these designs.

We collated BACI and R-BACI datasets by searching the Web of Science Core Collection (Thomson Reuters, 2019), which included the following citation indexes: Science Citation Index Expanded (SCI-EXPANDED) 1900-present; Social Sciences Citation Index (SSCI) 1900-present Arts & Humanities Citation Index (A&HCI) 1975-present; Conference Proceedings Citation Index - Science (CPCI-S) 1990-present; Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH) 1990-present; Book Citation Index - Science (BKCI-S) 2008-present; Book Citation Index - Social Sciences & Humanities (BKCI-SSH) 2008-present; Emerging Sources Citation Index (ESCI) 2015-present; Current Chemical Reactions (CCR-EXPANDED) 1985-present (Includes Institut National de la Propriete Industrielle structure data back to 1840); Index Chemicus (IC) 1993-present. The following search terms were used: ['BACI'] OR ['Before-After Control-Impact'] and the search was conducted on the 18th December 2017. Our search returned 674 results, which we then refined by selecting only 'Article' as the document type and using only the following Web of Science Categories: 'Ecology', 'Marine Freshwater Biology', 'Biodiversity Conservation', 'Fisheries', 'Oceanography', 'Forestry', 'Zoology', Ornithology', 'Biology', 'Plant Sciences',

‘Entomology’, ‘Remote Sensing’, ‘Toxicology’ and ‘Soil Science’. This left 579 results, which we then restricted to articles published since 2002 (15 years prior to search) to give us a realistic opportunity to obtain the raw datasets, thus reducing this number to 542. We were able to access the abstracts of 521 studies and excluded any that did not test the effect of an environmental intervention or threat using an R-BACI or BACI design with response measures related to the abundance (e.g., density, counts, biomass, cover), reproduction (reproductive success), or size (body length, body mass) of animals or plants. Many studies did not test a relevant metric (e.g., they measured species richness), did not use a BACI or R-BACI design, or did not test the effect of an intervention or threat — this left 96 studies for which we contacted all corresponding authors to ask for the raw dataset. We were able to fully access 54 raw datasets, but upon closer inspection we found that three of these datasets either: did not use a BACI design; did not use the metrics we specified; or did not provide sufficient data for our analyses. This left 51 datasets in total that we used in our preliminary analyses (Appendix S2).

All the datasets were originally collected to evaluate the effect of an environmental intervention or impact. Most of them contained multiple response variables (e.g., different measures for different species, such as abundance or density for species A, B, and C). Within a dataset, we use the term “response” to refer to the estimation of the causal effect on one response variable. There were 1,968 responses in total across 51 datasets. We then excluded 932 responses (resulting in the exclusion of one dataset) where one or more of the four time-period and treatment subsets (Before Control, Before Impact, After Control, and After Impact data) consisted of entirely zero measurements, or two or more of these subsets had more than 90% zero measurements. We also excluded one further dataset as it was the only one to not contain repeated measurements at sites in both the before- and after-periods. This was necessary to generate reliable standard errors when modelling these data. We modelled the remaining 1,036 responses from across 49 datasets (Table S1).

We applied each study design to the appropriate components of each dataset using Generalised Linear Models (GLMs; Bolker et al., 2009; Stroup, 2012) because of their generality and ability to implement the statistical estimators of many different study designs. The model structure of GLMs was adjusted for each response in each dataset based on the study design specified, response measure, and dataset structure (Table S2). We quantified the effect of the time period for the BA design (After vs Before the impact) and the effect of the treatment type for the CI and R-CI designs (Impact vs Control) on the response variable (Table S2). For BACI and R-BACI designs, we implemented two statistical estimators: 1.) a DiD estimator that estimated the true effect using an interaction term between time and

treatment type; and 2.) a covariance adjustment estimator that estimated the true effect using a term for the treatment type with a lagged variable (Table S2).

As there were large numbers of responses, we used general *a priori* rules to specify models for each response. These rules determined the error family of each GLM based on the nature of the measure used and preliminary data exploration as follows: count measures (e.g., abundance) = Poisson; density measures (e.g., biomass or abundance per unit area) = Quasipoisson, as data for these measures tended to be overdispersed; percentage measures (e.g., percentage cover) = Quasibinomial; and size measures (e.g., body length) = Gaussian. Whilst using general *a priori* rules may have led to some model misspecification, this is unlikely to have substantially affected our pairwise comparison of estimates from different designs.

We treated each year or season in which data were collected as independent observations because the implementation of a seasonal term in models is likely to vary on a case-by-case basis; this will depend on the research questions posed by each study and was not feasible for us to consider given the large number of responses we were modelling. The log link function was used for all models to generate a standardised log response ratio as an estimate of the true effect for each response; a fixed effect coefficient (a variable named treatment status; Table S2) was used to estimate the log response ratio (Gurevitch and Hedges, 1999). If the response had at least ten 'sites' (independent sampling units) and two measurements per site on average, we used the random effects of subsample (replicates within a site) nested within site to capture the dependence within a site and subsample (i.e., a Generalised Linear Mixed Model or GLMM (Bolker et al., 2009; Stroup, 2012) was implemented instead of a GLM); otherwise, we fitted a GLM with only the fixed effects (Table S2).

We fitted all models using R version 3.5.1 (R Core Team, 2019), and packages lme4 (Bates et al., 2015) and MASS (Venables and Ripley, 2002). Code to replicate all analyses is available from Zenodo: https://doi.org/10.5281/zenodo.3560856. We compared the estimates obtained using each study design (both in terms of point estimates and estimates with associated standard error) by their magnitude and sign.

A model-based quantification of the bias in study design estimates

We used a hierarchical Bayesian model motivated by the decomposition in Equation 1 to quantify the bias in different study design estimates. This model takes the estimated intervention effects and their standard errors as inputs. Let $\hat{\beta}_{ij}$ be the true effect estimator in study i using design j and $\hat{\sigma}_{ij}$ be its estimated standard error from the corresponding GLM or GLMM.

Our hierarchical model assumes:

$$\hat{\beta}_{ij} = \beta_i + \gamma_{ij} + \varepsilon_{ij},$$

$$\beta_i \sim \mathrm{N}(0, \sigma_\beta^2),\ \gamma_{ij} \sim \mathrm{N}(0, \sigma_j^2),\ \varepsilon_i \sim \mathrm{N}(0, \Lambda), \qquad \text{(Equation 2)}$$

where β_i is the true effect for response i, γ_{ij} is the bias of design j in response i, and $\varepsilon_{ij}$is the sampling noise of the statistical estimator. Although γ_{ij} technically incorporates both the design bias and any misspecification (modelling) bias due to using GLMs or GLMMs (Equation 1), we expect the modelling bias to be much smaller than the design bias (Light et al., 1990; Rubin, 2008). We assume the statistical errors ε_i within a response are related to the estimated standard errors through the following joint distribution:

$$\Lambda = \lambda \cdot \mathrm{diag}(\hat{\sigma}_i)\Omega\mathrm{diag}(\hat{\sigma}_i), \qquad \text{(Equation 3)}$$

where Ω is the correlation matrix for the different estimators in the same response and λ is a scaling factor to account for possible over/under-estimation of the standard errors.

This model effectively quantifies the bias of design j using the value of σ_j (larger values = more bias) by accounting for within-response correlations using the correlation matrix Ω and for possible under-estimation of the standard error using λ. We randomised the sign of $\hat{\beta}_{ij}$ as our model assumes that the bias of each design is randomly distributed across datasets and is on average zero. We ensured that the prior distributions we used had large variances so they would have a very small effect on the posterior distribution — accordingly we placed the following disperse priors on the variance parameters:

$$\sigma_\beta, \sigma_1, \ldots, \sigma_J \sim \text{Inv-Gamma}(1, 0.02),\ \lambda \sim \text{Gamma}(2,2),\ \Omega \sim \text{LKJ}(1) \qquad \text{(Equation 4)}.$$

We fitted the hierarchical Bayesian model in R version 3.5.1 using the Bayesian inference package rstan (Stan Development Team, 2020). All data and code analysed in this study are available from Zenodo: https://doi.org/10.5281/zenodo.3560856.

Results

Prevalence of study designs

We found that the biodiversity-conservation (Conservation Evidence) and social-science (Campbell Collaboration) literature had similarly high proportions of intervention studies that used CI designs and After designs, but low proportions of intervention studies that used R-BACI, BACI, or BA designs (Fig.2). There were slightly higher proportions of R-CI designs in social-science reviews than in the biodiversity-conservation literature (Fig.2). The R-BACI, R-CI, and BACI designs were used by 23% of studies on biodiversity conservation interventions, and 36% of studies on social interventions.

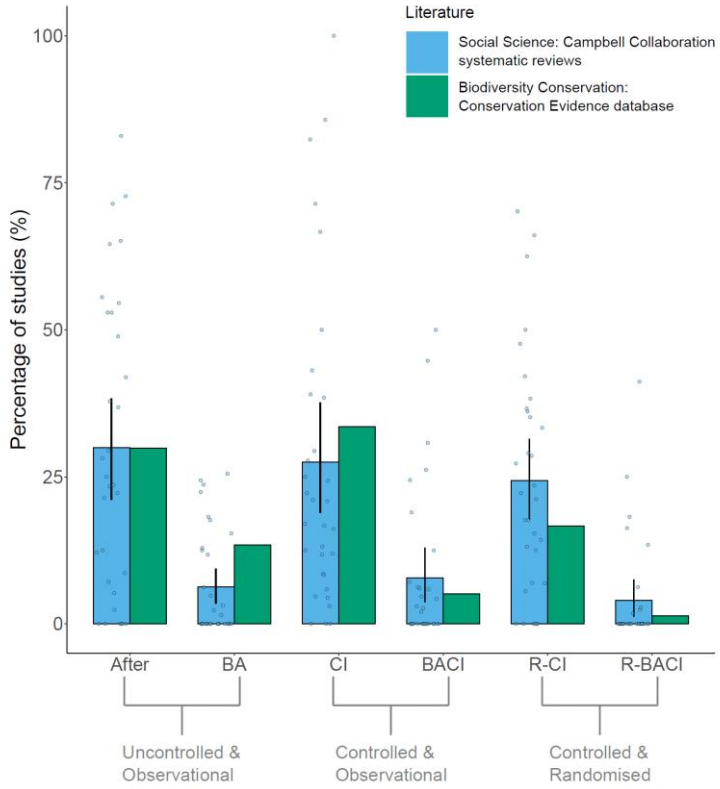


Figure 2 *– Percentage of studies with different study designs in the biodiversity-conservation and social-science literature. Studies from the biodiversity-conservation literature were screened from the Conservation Evidence database (n=4,260 studies) and studies from the social-science literature were screened from 32 Campbell Collaboration systematic reviews (n=1,009 studies – note studies excluded by these reviews based on their study design were still counted). Percentages for the social-science literature were calculated for each systematic review (blue data points) and then averaged across all 32 reviews (blue bars and black vertical lines represent mean and 95% Confidence Intervals, respectively). Percentages for the biodiversity-conservation literature are absolute values (green bars) calculated from the entire Conservation Evidence database (excluding reviews). BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact, R-BACI = Randomised BACI, R-CI = Randomised CI.*

Influence of different study designs on study results

In non-randomised datasets, we found that estimates of BACI (with covariance adjustment) and CI designs were very similar, whilst the point estimates for most other designs often differed substantially in their magnitude and sign. We found similar results in randomised datasets for R-BACI (with covariance adjustment) and R-CI designs. For approximately 30% of responses, in both non-randomised and randomised datasets, study design estimates differed in their statistical significance (i.e., $p<0.05$ versus $p>=0.05$), except for estimates of (R-)BACI (with covariance adjustment) and (R-)CI designs (Table 2; Fig.3). It was rare for the 95% confidence intervals of different designs' estimates to not overlap – except when comparing estimates of BA designs to (R-)BACI (with covariance adjustment) and (R-)CI designs (Table 2). It was even rarer for estimates of different designs to have significantly different signs (i.e., one estimate with entirely negative confidence intervals versus one with entirely positive confidence intervals; Table 2, Fig.3). Overall, point estimates often differed greatly in their magnitude and, to a lesser extent, in their sign between study designs, but did not differ as greatly when accounting for the uncertainty around point estimates – except in terms of their statistical significance.

Table 2 *– Pairwise comparison of estimates obtained using different study designs. This shows the proportion of responses in which there were differences in the magnitude (by >100%) and sign of estimates, and differences in the significance, sign and overlap between associated 95% confidence intervals. For randomised datasets, BACI and CI labels refer to R-BACI and R-CI designs (denoted by 'R-'). The 100% difference in magnitude criterion is set relative to the smaller estimate. DiD = Difference in Differences; CA = covariance adjustment. 95% Conf. Ints. refers to 95% Confidence Intervals and P.E. refers to point estimate. BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact.*

Randomised (R-)						
Design 1	**Design 2**	**No overlap (95% Conf. Ints.)**	**>100% difference in magnitude (P.E.)**	**Different significance (95% Conf. Ints.)**	**Different signs (P.E.)**	**Significantly different sign (95% Conf. Ints.)**
BACI DiD	BACI CA	0.01	0.68	0.27	0.32	0.00
BACI DiD	CI	0.01	0.69	0.27	0.32	0.00
BACI DiD	BA	0.01	0.68	0.29	0.34	0.00
BACI CA	CI	0.00	0.04	0.05	0.01	0.00
BACI CA	BA	0.16	0.82	0.33	0.47	0.06
CI	BA	0.16	0.82	0.30	0.47	0.07
Non-randomised						
Design 1	**Design 2**	**No overlap (95% Conf. Ints.)**	**>100% difference in magnitude (P.E.)**	**Different significance (95% Conf. Ints.)**	**Different signs (P.E.)**	**Significantly different sign (95% Conf. Ints.)**
BACI DiD	BACI CA	0.04	0.58	0.31	0.27	0.00
BACI DiD	CI	0.05	0.61	0.28	0.30	0.01
BACI DiD	BA	0.04	0.61	0.22	0.25	0.01
BACI CA	CI	0.00	0.18	0.08	0.08	0.00
BACI CA	BA	0.14	0.74	0.34	0.36	0.03
CI	BA	0.12	0.71	0.33	0.37	0.02

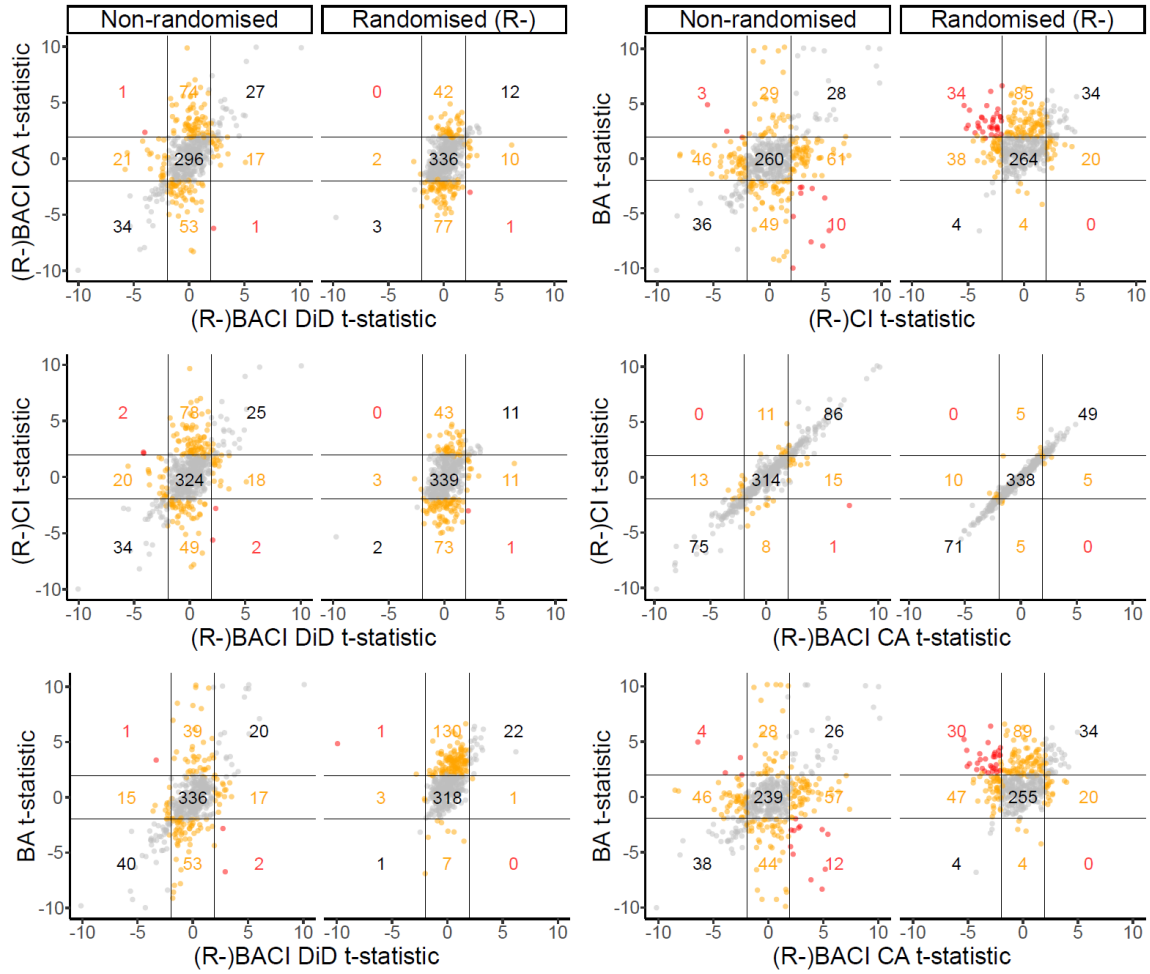


Figure 3 *– Pairwise comparisons of t-statistics for estimates obtained using different study designs for responses across 49 different datasets (non-randomised or randomised). t-statistics are obtained from two-sided t-tests of estimates obtained by each design for different responses in each dataset using Generalised Linear Models (see Materials and methods). For randomised datasets, BACI and CI axis labels refer to R-BACI and R-CI designs (denoted by 'R-'). DiD = Difference in Differences; CA = covariance adjustment. Lines at t-statistic values of 1.96 denote boundaries between cells and colours of points indicate differences in direction and statistical significance (p<0.05; grey = same sign and significance, orange = same sign but difference in significance, red = different sign and significance). Numbers refer to the number of responses in each cell. BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact.*

Levels of bias in estimates of different study designs

We modelled study design bias using a random effect across datasets in a hierarchical Bayesian model; σ is the standard deviation of the bias term, and assuming bias is randomly distributed across datasets and is on average zero, larger values of σ will indicate a greater magnitude of bias (see Materials and methods). We found that, for randomised datasets, estimates of both R-BACI (using covariance adjustment; CA) and R-CI designs were affected by negligible amounts of bias (very small values of σ; Table 3). When the R-BACI design used the DiD estimator, it suffered from slightly more bias (slightly larger values of σ), whereas the BA design had very high bias when applied to randomised datasets (very large values of σ; Table 3). There was a highly positive correlation between the estimates of R-BACI (using covariance adjustment) and R-CI designs (Ω[R-BACI CA, R-CI] was close to 1; Table 3). Estimates of R-BACI using the DiD estimator were also positively correlated with estimates of R-BACI using covariance adjustment and R-CI designs (moderate positive mean values of Ω[R-BACI CA, R-BACI DiD] and Ω[R-BACI DiD, R-CI]; Table 3).

For non-randomised datasets, controlled designs (BACI and CI) were substantially less biased (far smaller values of σ) than the uncontrolled BA design (Table 3). A BACI design using the DiD estimator was slightly less biased than the BACI design using covariance adjustment, which was, in turn, slightly less biased than the CI design (Table 3).

Standard errors estimated by the hierarchical Bayesian model were reasonably accurate for the randomised datasets (see λ in Materials and methods and Table 3), whereas there was some underestimation of standard errors and lack-of-fit for non-randomised datasets.

Table 3 *– Results of hierarchical Bayesian model for randomised and non-randomised datasets. In randomised datasets, BACI and CI terms refer to R-BACI and R-CI designs (denoted by 'R-'). The σ terms are the standard deviations of the bias of each design, so larger σ values correspond to more biased designs.* σ_β *refers to the standard deviation of the true effect across all datasets.* Ω *represents the within-response correlations between study design estimates, and* λ *models systematic underestimation (*$\lambda>1$*) or overestimation (*$\lambda<1$*) of the statistical error using GLM(M)s. See Materials and methods for more details. BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact.*

Randomised (R-)		
Term	**Posterior mean**	**95% Credible Interval**
σ_β	0.746	[0.679, 0.813]
λ	1.119	[0.980, 1.276]
σ[BACI DiD]	0.029	[0.005, 0.097]
σ[BACI CA]	0.005	[0.002, 0.008]
σ[CI]	0.005	[0.002, 0.008]
σ[BA]	0.773	[0.699, 0.846]
Ω[BACI DiD, BACI CA]	0.268	[0.152, 0.379]
Ω[BACI DiD, CI]	0.239	[0.122, 0.354]
Ω[BACI DiD, BA]	0.849	[0.770, 0.914]
Ω[BACI CA, CI]	0.995	[0.994, 0.996]
Ω[BACI CA, BA]	-0.168	[-0.332, 0.002]
Ω[CI, BA]	-0.184	[-0.349, -0.015]
Non-randomised		
Term	**Posterior mean**	**95% Credible Interval**
σ_β	0.700	[0.628, 0.776]
λ	1.822	[1.595, 2.098]
σ[BACI DiD]	0.017	[0.004, 0.049]
σ[BACI CA]	0.049	[0.005, 0.128]
σ[CI]	0.091	[0.008, 0.137]
σ[BA]	0.645	[0.573, 0.720]
Ω[BACI DiD, BACI CA]	0.140	[0.010, 0.263]
Ω[BACI DiD, CI]	0.036	[-0.106, 0.176]
Ω[BACI DiD, BA]	0.798	[0.718, 0.865]
Ω[BACI CA, CI]	0.939	[0.923, 0.954]
Ω[BACI CA, BA]	-0.127	[-0.285, 0.026]
Ω[CI, BA]	-0.229	[-0.397, -0.061]

Discussion

Our approach provides a principled way to quantify the levels of bias associated with different study designs. We found that randomised study designs (R-BACI and R-CI) and observational BACI designs are poorly represented in the environmental and social sciences; collectively, the After design (a post-impact time series), the uncontrolled BA design, and the observational CI design made up a substantially greater proportion of intervention studies (Fig.2). And yet R-BACI, R-CI, and BACI designs were found to be quantifiably less biased than other observational designs.

As expected, the R-CI and R-BACI designs (using a covariance adjustment estimator) performed well; the R-BACI design using a DiD estimator performed slightly less well, probably because the differencing of pre-impact data by this estimator may introduce additional statistical noise compared to covariance adjustment, which controls for these data using a lagged regression variable. Of the observational designs, the BA design performed very poorly (when analysing both randomised and non-randomised data) as expected, being uncontrolled and therefore prone to severe design bias (Christie et al., 2019; de Palma et al., 2018). The CI design also tended to be more biased than the BACI design (using a DiD estimator) due to pre-existing differences between the impact and control groups. For BACI designs, we recommend that the underlying assumptions of DiD and CA estimators are carefully considered before choosing to apply them to data collected for a specific research question (Angrist and Pischke, 2008; Ding and Li, 2019). Their levels of bias were negligibly different, and their known bracketing relationship suggests they will typically give estimates with the same sign, although their tendency to over- or underestimate the true effect will depend on how well the underlying assumptions of each are met (most notably, parallel trends for DiD and no unmeasured confounders for CA; see Introduction; Angrist and Pischke, 2008; Ding and Li, 2019). Overall, these findings demonstrate the power of large within-study comparisons to directly quantify differences in the levels of bias associated with different designs.

We must acknowledge that the assumptions of our hierarchical model (that the bias for each design (j) is on average zero and normally distributed) cannot be verified without gold standard randomised experiments and that, for observational designs, the model was overdispersed (potentially due to underestimation of statistical error by GLM(M)s or positively correlated design biases). The exact values of our hierarchical model should therefore be treated with appropriate caution, and future research is needed to refine and improve our approach to quantify these biases more precisely. Responses within datasets may also not be independent as multiple species could interact; therefore, the estimates analysed by our hierarchical model

are statistically dependent on each other, and although we tried to account for this using a correlation matrix (see Materials and methods, Equation 3), this is a limitation of our model. We must also recognise that we collated datasets using non-systematic searches (Gusenbauer and Haddaway, 2020; Konno and Pullin, 2020) and therefore our analysis potentially exaggerates the intrinsic biases of observational designs (i.e., our data may disproportionately reflect situations where the BACI design was chosen to account for confounding factors). We nevertheless show that researchers were wise to use the BACI design because it was less biased than CI and BA designs across a wide range of datasets from various environmental systems and locations. Without undertaking costly and time-consuming pre-impact sampling and pilot studies, researchers are also unlikely to know the levels of bias that could affect their results. Finally, we did not consider sample size, but it is likely that researchers might use larger sample sizes for CI and BA designs than BACI designs. This is, however, unlikely to affect our main conclusions because larger sample sizes could increase type I errors (false positive rate) by yielding more precise, but biased estimates of the true effect (Christie et al., 2019).

Our analyses provide several empirically supported recommendations for researchers designing future studies to assess an impact of interest. First, using a controlled and/or randomised design (if possible) was shown to strongly reduce the level of bias in study estimates. Second, when observational designs must be used (as randomisation is not feasible or too costly), we urge researchers to choose the BACI design over other observational designs — and when that is not possible, to choose the CI design over the uncontrolled BA design. Although we did not quantify the bias associated with the After design, we know from previous studies (Christie et al., 2019; de Palma et al. 2018) that this design suffers from both of the biases associated with BA and CI designs, and is only appropriate to describe or monitor rates of change over time after an impact has occurred. We acknowledge that limited resources, short funding timescales, and ethical or logistical constraints (Butsic et al., 2017) may force researchers to use the CI design (if randomisation and pre-impact sampling are impossible) or the BA design (if appropriate controls cannot be found; Christie et al., 2019). To facilitate the usage of less biased designs, longer-term investments in research effort and funding are required (França et al., 2016). Far greater emphasis on study designs in statistical education (Brownstein et al., 2019) and better training and collaboration between researchers, practitioners, and methodologists, is needed to improve the design of future studies. For example, we can potentially improve the CI design by pairing or matching the impact group and control group (Imbens and Rubin, 2015), or improve the BA design using regression discontinuity methods (Butsic et al., 2017; Hahn et al., 2001). Where the choice of study design is limited, researchers must transparently communicate the limitations and

uncertainty associated with their results. Researchers should also consider the costs involved in poor inferences resulting from the use of more biased study designs, including the costs of implementing ineffective or harmful decisions and interventions, as well as the potential costs to the credibility of scientific evidence from misinforming decision-makers (Wauchope, 2020, p127-128).

Our findings also have wider implications for evidence synthesis, specifically the exclusion of certain observational study designs from syntheses (the 'rubbish in, rubbish out' concept; Slavin, 1995, 1986). We believe that observational designs should be included in systematic reviews and meta-analyses, but that careful adjustments are needed to account for their potential biases. Exclusion of observational studies often results from subjective, checklist-based 'Risk of Bias' or quality assessments of studies (e.g., AMSTRAD 2 (Shea et al., 2017), ROBINS-I (Sterne et al., 2016), or GRADE (Guyatt et al., 2013)) that are not data-driven and often neglect to identify the actual direction, or quantify the magnitude, of possible bias introduced by observational studies when rating the quality of a review's recommendations. We also found that there was a small proportion of studies that used randomised designs (R-CI or R-BACI) or observational BACI designs (Fig.2), suggesting that systematic reviews and meta-analyses risk excluding a substantial proportion of the literature and limiting the scope of their recommendations if such exclusion criteria are used (Christie et al., 2020b; Davies and Gray, 2015; Lortie et al., 2015). This problem is compounded by the fact that, at least in conservation science, studies using randomised or BACI designs are strongly concentrated in Europe, Australasia, and North America (Christie et al., 2020a). Systematic reviews that rely on these few types of study designs are therefore likely to fail to provide decision-makers outside of these regions with locally relevant recommendations that they prefer (Gutzat and Dormann, 2020). The Covid-19 pandemic has highlighted the difficulties in making locally relevant evidence-based decisions using studies conducted in different countries with different demographics and cultures, and on patients of different ages, ethnicities, genetics, and underlying health issues (Greenhalgh, 2020). This problem is also acute for decision-makers working on biodiversity conservation in the tropical regions, where the need for conservation is arguably the greatest (i.e., where most of Earth's biodiversity exists; Barlow et al., 2018) but they either have to rely on very few well-designed studies that are not locally relevant (i.e., have low generalisability), or more studies that are locally relevant but less well-designed (Christie et al., 2020a, 2020b). Either option could lead decision-makers to take ineffective or inefficient decisions. In the long-term, improving the quality and coverage of scientific evidence and evidence syntheses across the world will help solve these issues, but shorter-term solutions to synthesising patchy evidence bases are required.

Our work furthers sorely needed research on how to combine evidence from studies that vary greatly in their design. Our approach is an alternative to conventional meta-analyses which tend to only weight studies by their sample size or the inverse of their variance (Gurevitch and Hedges, 1999); when studies vary greatly in their study design, simply weighting by inverse variance or sample size is unlikely to account for different levels of bias introduced by different study designs (see Equation 1). For example, a BA study could receive a larger weight if it had lower variance than a BACI study, despite our results suggesting a BA study usually suffers from greater design bias. Our model provides a principled way to weight studies by both the likely amount of bias introduced by their study design and their variance and is therefore a form of 'bias-adjusted meta-analysis' (Efthimiou et al., 2017; Rhodes et al., 2020; Stone et al., 2020; Turner et al., 2009; Welton et al., 2009). However, instead of relying on elicitation of subjective expert opinions on the bias of each study, we provide a data-driven, empirical quantification of study biases – an important step that was called for to improve such meta-analytic approaches (Turner et al., 2009; Welton et al., 2009).

Future research is needed to refine our methodology, but our empirically grounded form of bias-adjusted meta-analysis could be implemented as follows: 1.) collate studies for the same true effect, their effect size estimates, standard errors, and the type of study design; 2.) enter these data into our hierarchical model, where effect size estimates share the same intercept (the true effect), a random effect term due to design bias (whose variance is estimated by the method we used), and a random effect term for statistical noise (whose variance is estimated by the reported standard error of studies); 3.) fit this model and estimate the shared intercept/true effect. Heuristically, this can be thought of as weighting studies by both their design bias and their sampling variance and could be implemented on a dynamic meta-analysis platform (such as www.metadataset.com; Shackelford et al., 2021). This approach has substantial potential to develop evidence synthesis in fields (such as biodiversity conservation; Christie et al., 2020a, 2020b) with patchy evidence bases, where reliably synthesising findings from studies that vary greatly in their design is a fundamental challenge.

The utility of this approach is also important in the context of the exponential growth of scientific evidence bases (Bornmann and Mutz, 2015; Larsen and von Ins, 2010) and the need to design systems of collating, synthesising and critically appraising evidence that are more efficient (Marshall and Wallace, 2019; O'Connor et al., 2018; Thomas et al., 2017; Wallace et al., 2014). Automation of evidence assessment using a model-based approach such as ours could help to speed up evidence assessment; however, these approaches need to be rigorously tested further to ensure we maintain the high standards of rigour in evidence synthesis that

we strive to achieve (Marshall et al., 2020, 2015; Marshall and Wallace, 2019; O'Connor et al., 2018; Tsafnat et al., 2013; Wallace et al., 2014).

Our study has highlighted an often overlooked aspect of debates over scientific reproducibility: that the reliability of studies is fundamentally determined by study design. Testing the effectiveness of conservation and social interventions is undoubtedly of great importance given the current challenges facing biodiversity and society in general and the serious need for more evidence-based decision-making (Donnelly et al., 2018; Sutherland et al., 2004). And yet our findings suggest that quantifiably less biased study designs are poorly represented in the environmental and social sciences. Greater methodological training of researchers and funding for intervention studies, as well as stronger collaborations between methodologists and practitioners is needed to facilitate the use of less biased study designs. Better communication and reporting of the uncertainty associated with different study designs is also needed, as well as more meta-research (the study of research itself) to improve standards of study design (Ioannidis, 2018). Our hierarchical model provides a principled way to combine studies using a variety of study designs that vary greatly in their risk of bias, enabling us to make more efficient use of patchy evidence bases. Ultimately, we hope that researchers and practitioners testing interventions will think carefully about the types of study designs they use, and we encourage the evidence synthesis community to embrace alternative methods for combining evidence from heterogeneous sets of studies to improve evidence-based decision-making in all disciplines.

References

Altindag, O., Joyce, T.J., Reeder, J.A., 2019. Can Nonexperimental Methods Provide Unbiased Estimates of a Breastfeeding Intervention? A Within-Study Comparison of Peer Counseling in Oregon. Evaluation Review 43, 152–188. https://doi.org/10.1177/0193841X19865963

Angrist, J.D., Pischke, J.-S., 2008. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, New Jersey.

Barlow, J., França, F., Gardner, T.A., Hicks, C.C., Lennox, G.D., Berenguer, E., Castello, L., Economo, E.P., Ferreira, J., Guénard, B., Gontijo Leal, C., Isaac, V., Lees, A.C., Parr, C.L., Wilson, S.K., Young, P.J., Graham, N.A.J., 2018. The future of hyperdiverse tropical ecosystems. Nature 559, 517–526. https://doi.org/10.1038/s41586-018-0301-1

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models using lme4. Journal of Statistical Software 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Benson, K., Hartz, A.J., 2000. A Comparison of Observational Studies and Randomized, Controlled Trials. New England Journal of Medicine 342, 1878–1886. https://doi.org/10.1056/NEJM200006223422506

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24, 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bornmann, L., Mutz, R., 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66, 2215–2222. https://doi.org/https://doi.org/10.1002/asi.23329

Brownstein, N.C., Louis, T.A., O'Hagan, A., Pendergast, J., 2019. The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making. The American Statistician 73, 56–68. https://doi.org/10.1080/00031305.2018.1529623

Butsic, V., Lewis, D.J., Radeloff, V.C., Baumann, M., Kuemmerle, T., 2017. Quasi-experimental methods enable stronger inferences from observational data in ecology. Basic and Applied Ecology. https://doi.org/10.1016/j.baae.2017.01.005

Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N., Morris, R.E., 2018. The Internal And External Validity Of The Regression Discontinuity Design: A Meta-Analysis Of 15 Within-Study Comparisons. Journal of Policy Analysis and Management 37, 403–429. https://doi.org/10.1002/pam.22051

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020a. The challenge of biased evidence in conservation. Conservation Biology cobi.13577. https://doi.org/10.1111/cobi.13577

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020b. Poor availability of context-specific evidence hampers decision-making in conservation. Biological Conservation 248, 108666. https://doi.org/10.1016/j.biocon.2020.108666

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Cook, T.D., Shadish, W.R., Wong, V.C., 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management 27, 724–750. https://doi.org/10.1002/pam.20375

Davies, G.M., Gray, A., 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). Ecology and Evolution 5, 5295–5304. https://doi.org/10.1002/ece3.1782

de Palma, A., Sanchez-Ortiz, K., Martin, P.A., Chadwick, A., Gilbert, G., Bates, A.E., Börger, L., Contu, S., Hill, S.L.L., Purvis, A., 2018. Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity: Study Design, Time, Space and Synthesis, in: Bohan, D., Dumbrell, A., Woodward, G., Jackson, M. Next Generation Biomonitoring: Part 1. Elsevier, Oxford, pp. 163–199.

Dimick, J.B., Ryan, A.M., 2014. Methods for Evaluating Changes in Health Care Policy. JAMA 312, 2401. https://doi.org/10.1001/jama.2014.16153

Ding, P., Li, F., 2019. A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. Political Analysis 27, 605–615. https://doi.org/DOI: 10.1017/pan.2019.25

Donnelly, C.A., Boyd, I., Campbell, P., Craig, C., Vallance, P., Walport, M., Whitty, C.J.M., Woods, E., Wormald, C., 2018. Four principles to make evidence synthesis more useful for policy. Nature 558, 361–364. https://doi.org/10.1038/d41586-018-05414-4

dos Santos Ribas, L.G., Pressey, R.L., Loyola, R., Bini, L.M., 2020. A global comparative analysis of impact evaluation methods in estimating the effectiveness of protected areas. Biological Conservation 246, 108595. https://doi.org/10.1016/j.biocon.2020.108595

Duvendack, M., Hombrados, J.G., Palmer-Jones, R., Waddington, H., 2012. Assessing 'what works' in international development: meta-analysis for sophisticated dummies. Journal of Development Effectiveness 4, 456–471. https://doi.org/10.1080/19439342.2012.710642

Eddy, T.D., Pande, A., Gardner, J.P.A., 2014. Massive differential site-specific and species-specific responses of temperate reef fishes to marine reserve protection. Global Ecology and Conservation 1, 13–26. https://doi.org/10.1016/j.gecco.2014.07.004

Efthimiou, O., Mavridis, D., Debray, T.P.A., Samara, M., Belger, M., Siontis, G.C.M., Leucht, S., Salanti, G., 4, on behalf of G.W.P., 2017. Combining randomized and non-randomized evidence in network meta-analysis. Statistics in Medicine 36, 1210–1226. https://doi.org/10.1002/sim.7223

Fisher, R.A., 1925. Statistical methods for research workers, 1st ed. Oliver and Boyd, Edinburgh.

França, F., Louzada, J., Korasaki, V., Griffiths, H., Silveira, J.M., Barlow, J., 2016. Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. Journal of Applied Ecology 53, 1098–1105. https://doi.org/10.1111/1365-2664.12657

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer series in statistics, New York.

Geijzendorffer, I.R., van Teeffelen, A.J.A., Allison, H., Braun, D., Horgan, K., Iturrate-Garcia, M., Santos, M.J., Pellissier, L., Prieur-Richard, A.-H., Quatrini, S., 2017. How can global conventions for biodiversity and ecosystem services guide local conservation actions? Current Opinion in Environmental Sustainability 29, 145–150. https://doi.org/10.1016/j.cosust.2017.12.011

Goldenhar, L.M., Schulte, P.A., 1994. Intervention research in occupational health and safety. Journal of Occupational Medicine 36, 763–778. https://doi.org/10.1002/(SICI)1097-0274(199604)29:4%3C289::AID-AJIM2%3E3.0.CO;2-K

Greenhalgh, T., 2020. Will COVID-19 be evidence-based medicine's nemesis? PLOS Medicine 17, e1003266. https://doi.org/10.1371/journal.pmed.1003266

Greenhalgh, T., 2019. How to read a paper: the basics of Evidence Based Medicine, 6th ed. John Wiley & Sons, Ltd, Hoboken.

Gurevitch, J., Hedges, L. v., 1999. Statistical Issues in Ecological Meta-analyses. Ecology 80, 1142–1149. https://doi.org/10.1890/0012-9658(1999)080[1142:SIIEMA]2.0.CO;2

Gusenbauer, M., Haddaway, N.R., 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research Synthesis Methods 11, 181–217. https://doi.org/10.1002/jrsm.1378

Gutzat, F., Dormann, C.F., 2020. Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice. Environmental Management 66, 435–449. https://doi.org/10.1007/s00267-020-01312-6

Guyatt, G., Oxman, A.D., Sultan, S., Brozek, J., Glasziou, P., Alonso-Coello, P., Atkins, D., Kunz, R., Montori, V., Jaeschke, R., Rind, D., Dahm, P., Akl, E.A., Meerpohl, J., Vist, G., Berliner, E., Norris, S., Falck-Ytter, Y., Schünemann, H.J., 2013. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. Journal of Clinical Epidemiology 66, 151–157. https://doi.org/https://doi.org/10.1016/j.jclinepi.2012.01.006

Hahn, J., Todd, P., Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica 69, 201–209. https://www.jstor.org/stable/2692190

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge.

Ioannidis, J.P.A., 2018. Meta-research: Why research on research matters. PLOS Biology 16, e2005468. https://doi.org/10.1371/journal.pbio.2005468

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLOS Medicine 2, e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J.P.A., Contopoulos-Ioannidis, D.G., Haidich, A.B., Pappa, M., Pantazis, N., Kokori, S.I., Tektonidou, M.G., Contopoulos-Ioannidis, D.G., Ioannidis, J.P.A., Lau, J., 2001. Comparison of evidence of treatment effects in randomized and nonrandomized studies. Journal of the American Medical Association 286, 821–830. https://doi.org/10.1001/jama.286.7.821

John, L.K., Loewenstein, G., Prelec, D., 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23, 524–532. https://doi.org/10.1177%2F0956797611430953

Junker, J., Petrovan, S.O., Arroyo-RodrÍguez, V., Boonratana, R., Byler, D., Chapman, C.A., Chetry, D., Cheyne, S.M., Cornejo, F.M., CortÉs-Ortiz, L., Cowlishaw, G., Christie, A.P., Crockford, C., Torre, S.D.L., de Melo, F.R., Fan, P., Grueter, C.C., GuzmÁn-Caro, D.C., Heymann, E.W., Herbinger, I., Hoang, M.D., Horwich, R.H., Humle, T., Ikemeh, R.A., Imong, I.S., Jerusalinsky, L., Johnson, S.E., Kappeler, P.M., Kierulff, M.C.M., KonÉ, I., Kormos, R., Le, K.Q., Li, B., Marshall, A.J., Meijaard, E., Mittermeier, R.A., Muroyama, Y., Neugebauer, E., Orth, L., Palacios, E., Papworth, S.K., Plumptre, A.J., Rawson, B.M., Refisch, J., Ratsimbazafy, J., Roos, C., Setchell, J.M., Smith, R.K., Sop, T., Schwitzer, C., Slater, K., Strum, S.C., Sutherland, W.J., Talebi, M., Wallis, J., Wich, S., Williamson, E.A., Wittig, R.M., KÜhl, H.S., 2020. A Severe Lack of Evidence Limits Effective Conservation of the World's Primates. BioScience. https://doi.org/10.1093/biosci/biaa082

Kerr, N.L., 1998. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review 2, 196–217. https://doi.org/10.1207%2Fs15327957pspr0203_4

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Hutton, J., Altman, D.G., 2009. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLOS ONE 4. https://doi.org/10.1371/journal.pone.0007824

Konno, K., Pullin, A.S., 2020. Assessing the risk of bias in choice of search sources for environmental meta-analyses. Research Synthesis Methods 11, 698–713. https://doi.org/10.1002/jrsm.1433

LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review 604–620. https://www.jstor.org/stable/1806062

Larsen, P.O., von Ins, M., 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84, 575–603. https://doi.org/10.1007/s11192-010-0202-z

Light, R.J., Singer, J.D., Willett, J.B., 1990. By design: Planning research on higher education. Harvard University Press, Cambridge.

Long, Q., Little, R.J., Lin, X., 2008. Causal inference in hybrid intervention trials involving treatment choice. Journal of the American Statistical Association 103, 474–484.

Lortie, C.J., Stewart, G., Rothstein, H., Lau, J., 2015. How to critically read ecological meta-analyses. Research Synthesis Methods 6, 124–133. https://doi.org/10.1002/jrsm.1109

Marshall, I.J., Johnson, B.T., Wang, Z., Rajasekaran, S., Wallace, B.C., 2020. Semi-Automated evidence synthesis in health psychology: current methods and future prospects. Health Psychology Review 14, 145–158. https://doi.org/10.1080/17437199.2020.1716198

Marshall, I.J., Kuiper, J., Wallace, B.C., 2015. Automating Risk of Bias Assessment for Clinical Trials. IEEE Journal of Biomedical and Health Informatics 19, 1406–1412. https://doi.org/10.1109/JBHI.2015.2431314

Marshall, I.J., Wallace, B.C., 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews 8, 163. https://doi.org/10.1186/s13643-019-1074-9

McKinnon, M.C., Cheng, S.H., Garside, R., Masuda, Y.J., Miller, D.C., 2015. Sustainability: Map the evidence. Nature 528, 185–187. https://doi.org/10.1038/528185a

Moscoe, E., Bor, J., Bärnighausen, T., 2015. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. Journal of Clinical Epidemiology 68, 132–143. https://doi.org/https://doi.org/10.1016/j.jclinepi.2014.06.021

O'Connor, A.M., Tsafnat, G., Gilbert, S.B., Thayer, K.A., Wolfe, M.S., 2018. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic Reviews 7, 3. https://doi.org/10.1186/s13643-017-0667-4

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Peirce, C.S., Jastrow, J., 1884. On small differences in sensation. Memoirs of the National Academy of Sciences 3, 73-83.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rhodes, K.M., Savović, J., Elbers, R., Jones, H.E., Higgins, J.P.T., Sterne, J.A.C., Welton, N.J., Turner, R.M., 2020. Adjusting trial results for biases in meta-analysis: combining data-based evidence on bias with detailed trial assessment. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183, 193–209. https://doi.org/10.1111/rssa.12485

Rosenbaum, P.R., 2010. Design of observational studies. Springer, Cham. https://doi.org/10.1007/978-3-030-46405-9

Rubin, D.B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2, 808–840. https://doi.org/10.1214/08-AOAS187

Sagarin, R., Pauchard, A., 2010. Observational approaches in ecology open new ground in a changing world. Frontiers in Ecology and the Environment 8, 379–386. https://doi.org/10.1890/090001

Salmond, S.S., 2008. Randomized Controlled Trials: Methodological Concepts and Critique. Orthopaedic Nursing 27. https://doi.org/10.1097/01.NOR.0000315626.44137.94

Shackelford, G.E., Martin, P.A., Hood, A.S.C., Christie, A.P., Kulinskaya, E., Sutherland, W.J., 2021. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biology 19, 33. https://doi.org/10.1186/s12915-021-00974-w

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference, 1st ed. Houghton Mifflin, Boston.

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., Henry, D.A., 2017. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. British Medical Journal 358, 1–8. https://doi.org/10.1136/bmj.j4008

Sher, A.A., el Waer, H., González, E., Anderson, R., Henry, A.L., Biedron, R., Yue, P., 2018. Native species recovery after reduction of an invasive tree by biological control with and without active removal. Ecological Engineering 111, 167–175. https://doi.org/https://doi.org/10.1016/j.ecoleng.2017.11.018

Slavin, R.E., 1995. Best evidence synthesis: An intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48, 9–18. https://doi.org/10.1016/0895-4356(94)00097-A

Slavin, R.E., 1986. Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews. Educational Researcher 15, 5–11. https://doi.org/10.3102/0013189X015009005

Smokorowski, K.E., Randall, R.G., Canada, O., East, Q.S., Canada, O., Lakes, G., 2017. Cautions on using the Before-After-Control-Impact design in environmental effects monitoring programs. Facets 2, 212–232. https://doi.org/10.1139/facets-2016-0058

Stan Development Team, 2020. RStan: the R interface to Stan R package version 2.19.3.

Sterne, J.A.C., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.-W., Churchill, R.,

Deeks, J.J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F., Higgins, J.P.T., 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. British Medical Journal 355, i4919. https://doi.org/10.1136/bmj.i4919

Stewart-Oaten, A., Bence, J.R., 2001. Temporal and Spatial Variation in Environmental Impact Assessment. Ecological Monographs 71, 305–339. https://doi.org/10.1890/0012-9615(2001)071[0305:TASVIE]2.0.CO;2

Stone, J.C., Glass, K., Munn, Z., Tugwell, P., Doi, S.A.R., 2020. Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches. Journal of Clinical Epidemiology 117, 36–45. https://doi.org/10.1016/j.jclinepi.2019.09.010

Stroup, W.W., 2012. Generalized linear mixed models: modern concepts, methods and applications, 1st ed. CRC press, Boca Raton.

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19, 305–308. https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, Steven, Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., Albrecht, L., Horsley, T., Platt, J., Armstrong, R., Nguyen, P.H., Plovnick, R., Arno, A., Ivers, N., Quinn, G., Au, A., Johnston, R., Rada, G., Bagg, M., Jones, A., Ravaud, P., Boden, C., Kahale, L., Richter, B., Boisvert, I., Keshavarz, H., Ryan, R., Brandt, L., Kolakowsky-Hayner, S.A., Salama, D., Brazinova, A., Nagraj, S.K., Salanti, G., Buchbinder, R., Lasserson, T., Santaguida, L., Champion, C., Lawrence, R., Santesso, N., Chandler, J., Les, Z., Schünemann, H.J., Charidimou, A., Leucht, S., Shemilt, I., Chou, R., Low, N., Sherifali, D., Churchill, R., Maas, A., Siemieniuk, R., Cnossen, M.C., MacLehose, H., Simmonds, M., Cossi, M.-J., Macleod, M., Skoetz, N., Counotte, M., Marshall, I., Soares-Weiser, K., Craigie, S., Marshall, R., Srikanth, V., Dahm, P., Martin, N., Sullivan, K.,

Danilkewich, A., Martínez García, L., Synnot, A., Danko, K., Mavergames, C., Taylor, M., Donoghue, E., Maxwell, L.J., Thayer, K., Dressler, C., McAuley, J., Thomas, J., Egan, C., McDonald, Steve, Tritton, R., Elliott, J., McKenzie, J., Tsafnat, G., Elliott, S.A., Meerpohl, J., Tugwell, P., Etxeandia, I., Merner, B., Turgeon, A., Featherstone, R., Mondello, S., Turner, T., Foxlee, R., Morley, R., van Valkenhoef, G., Garner, P., Munafo, M., Vandvik, P., Gerrity, M., Munn, Z., Wallace, B., Glasziou, P., Murano, M., Wallace, S.A., Green, S., Newman, K., Watts, C., Grimshaw, J., Nieuwlaat, R., Weeks, L., Gurusamy, K., Nikolakopoulou, A., Weigl, A., Haddaway, N., Noel-Storr, A., Wells, G., Hartling, L., O'Connor, A., Wiercioch, W., Hayden, J., Page, M., Wolfenden, L., Helfand, M., Pahwa, M., Yepes Nuñez, J.J., Higgins, J., Pardo, J.P., Yost, J., Hill, S., Pearson, L., 2017. Living systematic reviews: 2. Combining human and machine effort. Journal of Clinical Epidemiology 91, 31–37. https://doi.org/10.1016/j.jclinepi.2017.08.011

Thomson Reuters, 2019. ISI Web of Knowledge [WWW Document]. URL http://www.isiwebofknowledge.com (accessed 12.18.17).

Tsafnat, G., Dunn, A., Glasziou, P., Coiera, E., 2013. The automation of systematic reviews. British Medical Journal 346, f139. https://doi.org/10.1136/bmj.f139

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S., Thompson, S.G., 2009. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172, 21–47. https://doi.org/10.1111/j.1467-985X.2008.00547.x

Underwood, A.J., 1991. Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations. Marine and Freshwater Research 42, 569–587. https://doi.org/10.1071/MF9910569

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth Edi. ed. Springer, New York.

Wallace, B.C., Dahabreh, I.J., Schmid, C.H., Lau, J., Trikalinos, T.A., 2014. Modernizing Evidence Synthesis for Evidence-Based Medicine, in: Greenes, R.A. (Second Edition (Ed.), Clinical Decision Support. Elsevier, Oxford, pp. 339–361. https://doi.org/10.1016/B978-0-12-398476-0.00012-9

Watson, M., Christoforou, P., Herrera, P., Preece, D., Carrell, J., Harmon, M., Krier, P., Lewis, S., Maiti, R., Skipper, W., Taylor, E., Walsh, J., Zalzalah, M., Alhadeff, L., Kempka, R., Lanigan, J., Lee, Z.S., White, B., Ishizaka, K., Lewis, R., Slatter, T., Dwyer-Joyce, R., Marshall, M., 2019. An analysis of the quality of experimental design and reliability of results in tribology research. Wear 426–427, 1712–1718. https://doi.org/10.1016/j.wear.2018.12.028

Wauchope, H.S., 2020. Working with large-scale population trend data in ecology and conservation: methods and applications. University of Cambridge. https://doi.org/10.17863/CAM.59354

Welton, N.J., Ades, A.E., Carlin, J.B., Altman, D.G., Sterne, J.A.C., 2009. Models for Potentially Biased Evidence in Meta-Analysis Using Empirically Based Priors. Journal of the Royal Statistical Society. Series A (Statistics in Society) 172, 119–136. https://doi.org/10.1111/j.1467-985X.2008.00548.x

Zhao, Q., Keele, L.J., Small, D.S., 2019. Comment: Will competition-winning methods for causal inference also succeed in practice? Statistical Science 34, 72–76. https://doi.org/10.1214/18-STS680

Supplementary Information

Figure S1

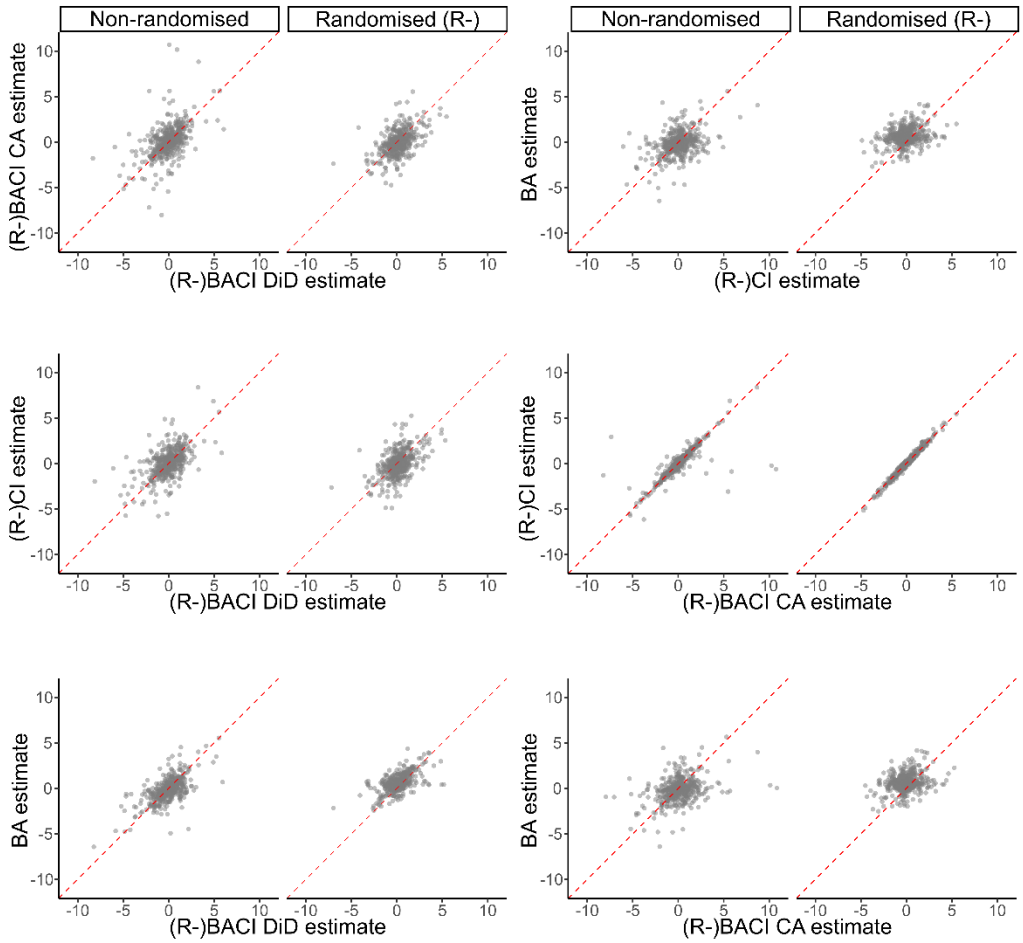


Figure S1 *– Pairwise comparisons of point estimates obtained using different study designs for 49 different datasets (non-randomised or randomised). For randomised datasets, BACI and CI axis labels refer to R-BACI and R-CI designs (denoted by 'R-'). DiD = Difference in Differences; CA = covariance adjustment. Red lines are a 1:1 line for visualising relationship. Two extreme outliers were removed to aid data visualisation of (non-randomised) BACI CA estimates. Source data are provided as a Source Data file. BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact.*

Table S1

Table S1 *– Summary of different datasets used in within-study comparison analyses, including: the unique identifier of each dataset, the number of responses extracted and modelled from each, the number of sites and subsamples, whether randomisation was used in the collection of the original data (Y = randomised, N = non-randomised), the response measures used, and whether the impact group and control group were compared within sites (Y = within site contrast, N = between site contrast).*

Dataset ID	No. responses	No. sites	No. subsamples	Randomised	Response measure	Within site contrast
1	2	6	6	N	Density	Y
2	1	5	5	N	Density	N
3	15	6	18	N	Count	N
4	57	6	6	Y	Density	Y
5	3	24	24	N	Density	N
6	3	2	20	N	Count	Y
7	5	3	108	N	Count	N
8	2	3	108	N	Count	N
9	4	3	9	N	Count, Density	N
10	1	8	61	N	Size	N
11	6	20	20	N	Density	Y
12	30	37	220	N	Count, Density	Y
13	1	4	4	N	Density	Y
14	53	26	26	N	Density	N
15	9	3	18	N	Density	Y
16	37	3	35	N	Count	N
17	2	2	382	N	Count	N
18	3	34	34	N	Count, Density	Y
19	28	6	6	N	Density	Y
20	4	2	2	N	Density	N
21	2	31	31	N	Percentage	N
22	30	28	28	N	Percentage	N
23	3	2	2	N	Density	N
24	50	35	35	N	Count	N
25	8	2	12	N	Density, Size, Count	Y
26	55	11	11	N	Density, Count	Y
27	21	1	1	N	Count	Y
28	1	18	18	N	Count	N
29	3	5	5	N	Density	Y
30	17	2	24	N	Count	Y
31	10	2	20	N	Count	Y
32	7	6	6	N	Count	N
33	2	8	32	Y	Count	Y
34	2	6	6	Y	Density	Y
35	13	4	467	N	Count	Y
36	1	3	3	N	Count	N
37	11	5	5	N	Count	N
38	2	4	4	N	Density	Y
39	18	6	6	N	Density	Y
40	29	3	3	N	Count	N
41	7	4	4	N	Count	Y
42	21	3	3	N	Count	Y
43	12	3	3	N	Count	Y

Dataset ID	No. responses	No. sites	No. subsamples	Randomised	Response measure	Within site contrast
44	10	4	40	N	Percentage	N
45	9	5	5	N	Density	Y
46	1	3	18	Y	Density	Y
47	2	3	3	N	Count, Size	Y
48	2	4	4	N	Count, Size	N
49	421	19	57	Y	Density	Y
Summary						
Datasets: 49	**Responses:** 1036			**Randomised:** 5 datasets **Non-randomised:** 44 datasets	**Count:** 27 datasets **Density:** 19 datasets **Percentage:** 3 datasets **Size:** 4 datasets	**Within-site contrast:** 27 datasets **Between-site contrast:** 22 datasets

Table S2

Table S2 *– Information on how different designs, and statistical methods therein, were applied to different subsets of each dataset using Generalised Linear (Mixed) Models (GL(M)Ms). DiD = Difference in Differences, CA = Covariance Adjustment. Response refers to the value of the response measure; treatment type refers to the impact or control group; time refers to the time period (before or after the impact occurred); treatment status refers to whether the site was subjected to the impact in that time period. BA = Before-After, CI = Control-Impact, BACI = Before-After-Control-Impact, R-BACI = Randomised BACI, R-CI = Randomised CI.*

Study design	**Statistical method**	**Subset of dataset used**	**Fixed Effects Model structure**
BACI R-BACI	DiD	All data	Response ~ treatment type + time + treatment status
	CA	All data	Post-impact within-site average ~ treatment status + pre-impact within-site average
CI R-CI	Difference	Data collected after impact (time = After)	Response ~ treatment status
BA	Difference	Impact data (treatment type = Impact)	Response ~ treatment status

Appendix S1

Reference	Campbell Collaboration Coordinating Group
Barlow, J., Bennett, C., Midgley, N., Larkin, S.K., Wei, Y., 2015. Parent-infant Psychotherapy for Improving Parental and Infant Mental Health: A Systematic Review. Campbell Systematic Reviews 11, 1-223. https://doi.org/10.4073/csr.2015.6	Social Welfare
Brody, C. et al., 2015. Economic Self-Help group Programs for Improving Women's Empowerment: A Systematic Review. Campbell Systematic Reviews 11, 1-182. https://doi.org/10.4073/csr.2015.19	International Development
Coren, E., Hossain, R., Pardo, J.P., Bakker, B., 2016. Interventions for promoting reintegration and reducing harmful behaviour and lifestyles in street-connected children and young people: a systematic review. Campbell Systematic Reviews 12, 1-198. https://doi.org/10.4073/csr.2016.5	International Development
De La Rue, L., Polanin, J.R., Espelage, D.L., Pigott, T.D., 2014. School-Based Interventions to Reduce Dating and Sexual Violence: A Systematic Review. Campbell Systematic Reviews 10, 1-110. https://doi.org/10.4073/csr.2014.7	Education
Fellmeth, G.L., Heffernan, C., Nurse, J., Habibula, S., Sethi, D., 2013. Educational and Skills-Based Interventions for Preventing Relationship and Dating Violence in Adolescents and Young Adults: A Systematic Review. Campbell Systematic Reviews 9, 1-124. https://doi.org/10.4073/csr.2013.14	Social Welfare
Filges, T., Geerdsen, L.P., Knudsen, A.D., Jørgensen, A.K., Kowalski, K., 2013. Unemployment Benefit Exhaustion: Incentive Effects on Job Finding Rates: A Systematic Review. Campbell Systematic Reviews 9, 1-104. https://doi.org/10.4073/csr.2013.4	Social Welfare
Filges, T., Montgomery, E., Kastrup, M., Jørgensen, A.K., 2015. The Impact of Detention on the Health of Asylum Seekers: A Systematic Review. Campbell Systematic Reviews 11, 1-104. https://doi.org/10.4073/csr.2015.13	Social Welfare
Filges, T., Smedslund, G., Knudsen, A.D., Jørgensen, A.K., 2015. Active Labour Market Programme Participation for Unemployment Insurance Recipients: A Systematic Review. Campbell Systematic Reviews 11, 1-342. https://doi.org/10.4073/csr.2015.2	Social Welfare
Fleming, P. et al., 2019. Individualized funding interventions to improve health and social care outcomes for people with a disability: A mixed-methods systematic review. Campbell Systematic Reviews 15, e1008. https://doi.org/10.4073/csr.2019.3	Disability
Fong, C.J., Murphy, K.M., Westbrook, J.D., Markle, M.M., 2015. Behavioral, Psychological, Educational, and Vocational Interventions to Facilitate Employment Outcomes for Cancer Survivors: A Systematic Review. Campbell Systematic Reviews 11, 1-81. https://doi.org/10.4073/csr.2015.5	Education
Graham, C.W., West, M.D., Bourdon, J.L., Inge, K.J., Seward, H.E., 2016. Employment Interventions for Return to Work in Working Aged Adults Following Traumatic Brain Injury (TBI): A Systematic Review. Campbell Systematic Reviews 12, 1-133. https://doi.org/10.4073/csr.2016.6	Education

Reference	Campbell Collaboration Coordinating Group
Harada, T., Tsutomi, H., Mori, R., Wilson, D.B., 2019. Cognitive-behavioural treatment for amphetamine-type stimulants (ATS)-use disorders. Campbell Systematic Reviews 15, e1026. https://doi.org/10.1002/cl2.1026	Crime and Justice
Iemmi, V. et al., 2015. Community-based Rehabilitation for People With Disabilities in Low- and Middle-income Countries: A Systematic Review. Campbell Systematic Reviews 11, 1-177. https://doi.org/10.4073/csr.2015.15	International Development
Kristjansson, E. et al., 2015. Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically Disadvantaged Children Aged Three Months to Five Years: A Systematic Review. Campbell Systematic Reviews 11, 1-226. https://doi.org/10.4073/csr.2015.11	International Development, Nutrition, Social Welfare
Lindstrøm, M. et al., 2015. Family Behavior Therapy (FBT) for Young People in Treatment for Non-opioid Drug Use: A Systematic Review. Campbell Systematic Reviews 11, 1-77. https://doi.org/10.4073/csr.2015.9	Social Welfare
Lindstrøm, M. et al., 2015. Family Behavior Therapy (FBT) for Young People in Treatment for Non-opioid Drug Use: A Systematic Review. Campbell Systematic Reviews 11, 1-77. https://doi.org/10.4073/csr.2015.9	Social Welfare
Maynard, B.R. et al., 2015. Psychosocial Interventions for School Refusal with Primary and Secondary School Students: A Systematic Review. Campbell Systematic Reviews 11, 1-76. https://doi.org/10.4073/csr.2015.12	Education
Maynard, B.R., Solis, M.R., Miller, V.L., Brendel, K.E., 2017. Mindfulness-based interventions for improving cognition, academic achievement, behavior, and socioemotional functioning of primary and secondary school students. Campbell Systematic Reviews 13, 1-144. https://doi.org/10.4073/CSR.2017.5	Education, Social Welfare
Parker, B., Turner, W., 2013. Psychoanalytic/Psycho-dynamic Psychotherapy for Children and Adolescents Who Have Been Sexually Abused: A Systematic Review. Campbell Systematic Reviews 9, 1-58. https://doi.org/10.4073/csr.2013.13	Social Welfare
Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M.E., Lavenberg, J.G., 2013. Scared Straight and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency: A Systematic Review. Campbell Systematic Reviews 9, 1-55. https://doi.org/10.4073/csr.2013.5	Crime and Justice
Polec, L.A. et al., 2015. Strategies to Increase the Ownership and Use of Insecticide-Treated Bednets to Prevent Malaria. Campbell Systematic Reviews 11, 1-127. https://doi.org/10.4073/csr.2015.17	International Development
Regehr, C., Alaggia, R., Dennis, J., Pitts, A., Saini, M., 2013. Interventions to Reduce Distress in Adult Victims of Sexual Violence and Rape: A Systematic Review. Campbell Systematic Reviews 9, 1-133. https://doi.org/10.4073/csr.2013.3	Social Welfare
Reichow, B., Barton, E.E., Boyd, B.A., Hume, K., 2014. Early Intensive Behavioral Intervention (EIBI) for Young Children with Autism Spectrum Disorders (ASD): A Systematic Review. Campbell Systematic Reviews 10, 1-116. https://doi.org/10.4073/csr.2014.9	Education

Reference	Campbell Collaboration Coordinating Group
Rivas, C. et al., 2016. Advocacy Interventions to Reduce or Eliminate Violence and Promote the Physical and Psychosocial Well-Being of Women who Experience Intimate Partner Abuse: A Systematic Review. Campbell Systematic Reviews 12, 1-202. https://doi.org/10.4073/csr.2016.2	Social Welfare
Rohwer, A., Motaze, N.V., Rehfuess, E., Young, T., 2017. E-learning of evidence-based health care (EBHC) to increase EBHC competencies in healthcare professionals: a systematic review. Campbell Systematic Reviews 13, 1-147. https://doi.org/10.4073/csr.2017.4	Education
Samii, C. et al., 2014. Effects of Payment for Environmental Services (PES) on Deforestation and Poverty in Low and Middle Income Countries: A Systematic Review. Campbell Systematic Reviews 10, 1-95. https://doi.org/10.4073/csr.2014.11	International Development
Spier, E. et al., 2016. Parental, Community, and Familial Support Interventions to Improve Children's Literacy in Developing Countries: A Systematic Review. Campbell Systematic Reviews 12, 1-98. https://doi.org/10.4073/csr.2016.4	Education, International Development
Thomson, H., Thomas, S., Sellström, E., Petticrew, M., 2013. Housing Improvements for Health and Associated Socio-Economic Outcomes: A Systematic Review. Campbell Systematic Reviews 9, 1-348. https://doi.org/10.4073/csr.2013.2	Social Welfare
Toon, C., Gurusamy, K., 2014. Forensic Nurse Examiners versus Doctors for the Forensic Examination of Rape and Sexual Assault Complainants: A Systematic Review. Campbell Systematic Reviews 10, 1-56. https://doi.org/10.4073/csr.2014.5	Crime and Justice
Turner, H., Ncube, M., Turner, A., Boruch, R., Ibekwe, N., 2018. What are the effects of Teach For America on Math, English Language Arts, and Science outcomes of K–12 students in the USA?. Campbell Systematic Reviews 14, 1-60. https://doi.org/10.4073/csr.2018.7	Education
Walsh, K., Zwi, K., Woolfenden, S., Shlonsky, A., 2015. School-based Education Programmes for the Prevention of Child Sexual Abuse: A Systematic Review. Campbell Systematic Reviews 11, 1-180. https://doi.org/10.4073/csr.2015.10	Social Welfare
Welch, V.A. et al., 2016. Deworming and adjuvant interventions for improving the developmental health and well-being of children in low- and middle-income countries: a systematic review and network meta-analysis. Campbell Systematic Reviews 12, 1-383. https://doi.org/10.4073/csr.2016.7	International Development

Appendix S2

Reference
Abecasis, D., Afonso, P., O'Dor, R.K., Erzini, K., 2013. Small MPAs do not protect cuttlefish (Sepia officinalis). Fisheries Research 147, 196–201.
Adjeroud, M. et al., 2016. Localised and limited impact of a dredging operation on coral cover in the northwestern lagoon of New Caledonia. Marine Pollution Bulletin 105, 208–214.
Antón, A., Elosegi, A., García-Arberas, L., Díez, J., Rallo, A., 2011. Restoration of dead wood in Basque stream channels: effects on brown trout population. Ecology of Freshwater Fish 20, 461–471.
Baldigo, B., Warren, D., Ernst, A., Mulvihill, C., 2008. Response of Fish Populations to Natural Channel Design Restoration in Streams of the Catskill Mountains, New York. North American Journal of Fisheries Management 28, 954–969.
Barrientos, R. et al., 2012. Wire marking results in a small but significant reduction in avian mortality at power lines: A BACI designed study. PLOS ONE 7.
Barros, Á., Álvarez, D., Velando, A., 2014. Long-term reproductive impairment in a seabird after the Prestige oil spill. Biology Letters 10, 20131041.
Battershill, C.N. et al., 1993. A survey of the marine habitats and communities of Kapiti Island. Report. New Zealand Oceanographic Institute, Wellington, New Zealand.
Bicknell, J.E., Struebig, M.J., Davies, Z.G., 2015. Reconciling timber extraction with biodiversity conservation in tropical forests using reduced-impact logging. Journal of Applied Ecology 52, 379–388.
Burge, O.R. et al., 2017. Glyphosate redirects wetland vegetation trajectory following willow invasion. Applied Vegetation Science 20, 620–630.
Ceia, R.S., Machado, R.A., Ramos, J.A., 2016. Nestling food of three hole-nesting passerine species and experimental increase in their densities in Mediterranean oak woodlands. European Journal of Forest Research 135, 839–847.
Cibils, L., Principe, R., Gari, N., 2013. Effect of a dam on epilithic algal communities of a mountain stream: Before-after dam construction comparison. Journal of limnology 72, 79–94.
Clarke, S., Tully, O., 2014. BACI monitoring of effects of hydraulic dredging for cockles on intertidal benthic habitats of Dundalk Bay, Ireland. Journal of the Marine Biological Association of the United Kingdom 94, 1451–1464.
Claudet, J., Pelletier, D., Jouvenel, J.-Y., Bachet, F., Galzin, R. 2006. Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: Identifying community-based indicators. Biological Conservation 130, 349–369.
Craig, M.D. 2007. The short-term effects of edges created by forestry operations on the bird community of the jarrah forest, south-western Australia. Austral Ecology 32, 386–396.
Dietl, G.P., Durham, S.R. 2016. Geohistorical records indicate no impact of the Deepwater Horizon oil spill on oyster body size. Royal Society Open Science 3, 160763.
Donovan, M.K. et al., 2016. Effects of gear restriction on the abundance of juvenile fishes along sandy beaches in Hawai'i. PLOS ONE 11, e0155221.
Eddy, T.D., Pande, A., Gardner, J.P.A. 2014. Massive differential site-specific and species-specific responses of temperate reef fishes to marine reserve protection. Global Ecology and Conservation 1, 13–26.
Edwards, P.M., Shaloum, G., Bedell, D., 2018. A unique role for citizen science in ecological restoration: a case study in streams. Restoration Ecology 26, 29–35.

Reference
França, F. et al., 2016. Do space-for-time assessments underestimate the impacts of logging on tropical biodiversity? An Amazonian case study using dung beetles. Journal of Applied Ecology 53, 1098–1105.
França, F., Barlow, J., Araújo, B., Louzada, J., 2016. Does selective logging stress tropical forest invertebrates? Using fat stores to examine sublethal responses in dung beetles. Ecology and Evolution 6, 8526–8533.
Kelaher, B.P. et al., 2015. Strengthened enforcement enhances marine sanctuary performance. Global Ecology and Conservation 3, 503–510.
Major, H.L., Buxton, R.T., Schacter, C.R., Conners, M.G., Jones, I.L., 2017. Habitat modification as a means of restoring crested auklet colonies. The Journal of Wildlife Management 81, 112–121.
Mateos-Molina, D., Schärer-Umpierre, M.T., Appeldoorn, R.S., García-Charton, J.A., 2014. Measuring the effectiveness of a Caribbean oceanic island no-take zone with an asymmetrical BACI approach. Fisheries Research 150, 1–10.
McConnaughey, R.A., Syrjala, S.E. 2014. Short-term effects of bottom trawling and a storm event on soft-bottom benthos in the eastern Bering Sea. ICES Journal of Marine Science 71, 2469–2483.
Meroni, M. et al., 2017. Remote sensing monitoring of land restoration interventions in semi-arid environments with a before–after control-impact statistical design. International Journal of Applied Earth Observation and Geoinformation 59, 42–52.
Mills, K., Hamer, P., Quinn, G.P., 2017. Artificial reefs create distinct assemblages: a study of fish assemblage response to the deployment of artificial patch reefs. Marine Ecology Progress Series 585.
Moland, E. et al., 2013. Lobster and cod benefit from small-scale northern marine protected areas: inference from an empirical before–after control-impact study. Proceedings of the Royal Society B: Biological Sciences 280, 20122679.
Montefalcone, M. et al., 2008. BACI design reveals the decline of the seagrass Posidonia oceanica induced by anchoring. Marine Pollution Bulletin 56, 1637–1645.
Noreika, N. et al., 2016, Specialist butterflies benefit most from the ecological restoration of mires. Biological Conservation 196, 103–114.
Núñez, S.F. et al., 2019. Echolocation and stratum preference: key trait correlates of vulnerability of insectivorous bats to tropical forest fragmentation. Frontiers in Ecology and Evolution 7, 373.
Pande, A., Gardner, J.P.A., 2012. The Kapiti Marine Reserve (New Zealand): spatial and temporal comparisons of multi-species responses after 8 years of protection. New Zealand Journal of Marine and Freshwater Research 46, 71–89.
Pitcher, C.R., Burridge, C.Y., Wassenberg, T.J., Hill, B.J., Poiner, I.R., 2009. A large scale BACI experiment to test the effects of prawn trawling on seabed biota in a closed area of the Great Barrier Reef Marine Park, Australia. Fisheries Research 99, 168–183.
Rinella, M.J., Dean, R., Vavra, M., Parks, C.G., 2012. Vegetation responses to supplemental winter feeding of elk in western Wyoming. Western North American Naturalist 72, 78–83.
Schmitter-Soto, J.J. et al., 2018. Interdecadal trends in composition, density, size, and mean trophic level of fish species and guilds before and after coastal development in the Mexican Caribbean. Biodiversity and conservation 27, 459–474.
Schwerk, A., Dymitryszyn, I., 2017. Mowing intensity influences degree of changes in carabid beetle assemblages. Applied Ecology and Environmental Research 15, 427–440.
Sepúlveda, R.D., Valdivia, N., 2016. Localised effects of a mega-disturbance: spatiotemporal responses of intertidal sandy shore communities to the 2010 Chilean earthquake. PLOS ONE 11, e0157910.

Reference
Shaffer, J.A., Buhl, D.A., 2016. Effects of wind-energy facilities on breeding grassland bird distributions. Conservation Biology 30, 59–71.
Sharma, S. et al., 2016. Do restored oyster reefs benefit seagrasses? An experimental study in the Northern Gulf of Mexico. Restoration Ecology 24, 306–313.
Stagnol, D., Renaud, M., Davoult, D., 2013. Effects of commercial harvesting of intertidal macroalgae on ecosystem biodiversity and functioning. Estuarine, Coastal and Shelf Science 130, 99–110.
Stanley, T.R., Knopf, F.L., 2002. Avian responses to late-season grazing in a shrub-willow floodplain. Conservation Biology 16, 225–231.
Stokesbury, K.D.E., Harris, B., 2006. Impact of limited short-term sea scallop fishery on epibenthic community of Georges Bank closed areas. Marine Ecology Progress Series 307, 85–100.
Torres, A., Palacín, C., Seoane, J., Alonso, J.C., 2011. Assessing the effects of a highway on a threatened species using Before–During–After and Before–During–After-Control–Impact designs. Biological Conservation 144, 2223–2232.
van Deurs, M. et al., 2012. Short-and long-term effects of an offshore wind farm on three species of sandeel and their sand habitat. Marine Ecology Progress Series 458, 169–180.
Vandendriessche, S., Derweduwen, J., Hostens, K., 2015. Equivocal effects of offshore wind farms in Belgium on soft substrate epibenthos and fish assemblages. Hydrobiologia 756, 19–35.
Vehanen, T. et al., 2010. APPLIED ISSUES: Effects of habitat rehabilitation on brown trout (Salmo trutta) in boreal forest streams. Freshwater Biology 55, 2200–2214.
Vieira, J.V. et al., 2016. Assessment the short-term effects of wrack removal on supralittoral arthropods using the M-BACI design on Atlantic sandy beaches of Brazil and Spain. Marine Environmental Research 119, 222–237.
Watts, C., Armstrong, D., Innes, J., Thornburrow, D., 2011. Dramatic increases in weta (Orthoptera) following mammal eradication on Maungatautari - evidence from pitfalls and tracking tunnels. New Zealand Journal of Ecology 35.
Welsh, H.H., Waters, J.R., Hodgson, G.R., Weller, T.J., Zabel, C.J., 2015. Responses of the woodland salamander Ensatina eschscholtzii to commercial thinning by helicopter in late-seral Douglas-fir forest in northwest California. Forest Ecology and Management 335, 156–165.
Williams, D.E., Miller, M.W., Bright, A.J., Cameron, C.M., 2014. Removal of corallivorous snails as a proactive tool for the conservation of acroporid corals. PeerJ 2, e680.

4 | The challenge of biased evidence in conservation

This chapter was published as:

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020. The challenge of biased evidence in conservation. Conservation Biology cobi.13577. https://doi.org/10.1111/cobi.13577

Abstract

Conservation efforts to tackle the current biodiversity crisis need to be as efficient and effective as possible given the chronic underfunding of conservation. To inform decision-makers of the most effective conservation actions, it is important to identify biases and gaps in the conservation literature to prioritise future evidence generation. We assessed the state of this global literature base on amphibians and birds using the Conservation Evidence database, a comprehensive collection of quantitative tests of conservation actions (interventions) from the published literature. We investigated the spatial and taxonomic spread of studies from this database, as well as the distribution across biomes, effectiveness metrics, and study designs for these two taxonomic groups. Studies were heavily concentrated in Western Europe and North America for birds and particularly for amphibians, whilst temperate forest and grassland biomes were highly represented relative to the percentage of the Earth's terrestrial area they covered. Studies that used the most reliable study designs – Before-After Control-Impact and Randomised Control-Impact (or Randomised Controlled Trials) – were the most geographically restricted and scarce in the evidence base. Furthermore, there were negative spatial relationships between the numbers of studies and the numbers of threatened and data-deficient species across the world. Taxonomic biases and gaps were apparent for amphibians and birds – some entire orders were absent from the evidence base, whilst others were poorly represented relative to the proportion of threatened species they contained. The metrics used to evaluate the effectiveness of a given conservation action were often inconsistent between studies, potentially making them less directly comparable, and evidence synthesis more difficult. Future research should prioritise testing conservation actions on threatened species outside of Western Europe, North America, and Australasia. Standardising metrics and improving the rigor of study designs used to test conservation actions would also improve the quality of the evidence base for synthesis and decision-making.

Introduction

Biodiversity conservation receives insufficient funding to effectively combat the biodiversity crisis (Dirzo et al., 2014). This means that conservation researchers and funders must prioritise research effort to maximise its potential to inform conservation efforts. Whilst evidence-based conservation is ultimately likely to lead to more efficient conservation efforts, this approach requires a reliable evidence base. Efforts to summarise the evidence in conservation relating to the effectiveness of different conservation actions ('interventions'; Sutherland et al., 2004) have produced a substantial evidence base (Sutherland et al., 2019), yet little is known about the biases and gaps in this evidence. Characterising the current state of the evidence base for conservation is crucial to prioritising future research efforts (Aranda et al., 2011). In this paper, we focus on studies that test conservation interventions, such as restoring grasslands for birds or creating ponds for amphibians.

The lack of resources in conservation research are likely to lead to several forms of bias in the evidence base for conservation. Such biases may limit our ability to provide relevant evidence-based recommendations to decision-makers or make the process of evidence synthesis more challenging. For example, geographical and taxonomic biases towards certain regions or groups (e.g., wealthier countries or charismatic species) may lead to little evidence being available for certain local contexts. Alternatively, bias could be useful if research effort is prioritised to where it is needed most in conservation – for example, focusing the majority of studies on threatened species. Wealthier countries (e.g., those in North America, Europe, and Australia) perform the majority of conservation research and so we may expect that patterns of evidence will follow physical proximity to these countries (Reddy and Dávalos, 2003), as well as various socio-economic variables (e.g., GDP per capita, affluence, language, security, conflict, and infrastructure; Amano and Sutherland, 2013; Hickisch et al., 2019; Martin et al., 2012; Meyer et al., 2016). These factors are likely to cause publication bias (i.e., underrepresentation of studies from non-English speaking countries; Amano et al., 2016; Nuñez et al., 2019) and affect the representation of different habitats in the evidence base (Fazey et al., 2005). Research effort is also known to vary with taxonomic group (Clark and May, 2002; Donaldson et al., 2016; Murray et al., 2015), and to depend on the range size, diet, and body size of species (Brooke et al., 2014), contributing to biases towards larger, more detectable species (e.g., Brodie, 2009; Cardoso et al., 2011). These forms of bias affect the external validity of studies in the evidence base and are therefore important to help us understand how much evidence is available to inform conservation in different contexts.

Other forms of bias may also complicate the synthesis of evidence. Bias in the usage of different types of metrics to assess the effectiveness of the same conservation intervention

may make approaches such as meta-analyses difficult to use. This is because studies are less directly comparable if they use different types of metrics to assess effectiveness, thus reducing the number of studies that can be combined in a meta-analysis. For example, it would be difficult to combine a set of studies measuring reproductive success, reductions in adult mortality, numbers of individuals, and species richness of birds using nest boxes in a conventional meta-analysis on the effectiveness of nest boxes. Clearly, different metrics may be useful to assess different aspects of an intervention's effectiveness, as well as to give greater confidence about the overall effectiveness of an intervention. However, wide variation in metrics used to test the same intervention could cause confusion for decision-makers, especially if studies using different metrics yield different results (Capmourteres and Anand, 2016).

Differences in study quality due to different study designs may also make it more difficult to decide which studies to trust over others, particularly if they give conflicting results. Several different study designs are used to assess impacts of threats and interventions in ecology (Christie et al., 2019; de Palma et al., 2018), all of which are affected by different sources and levels of bias and noise. These range from relatively reliable designs such as experimental Randomised Control-Impact (RCI) (also known as Randomised Controlled Trials or RCTs) and quasi-experimental Before-After Control-Impact designs (BACI), to less reliable designs such as Control-Impact (CI), Before-After (BA) and After (Table 1). Evidence may also come in the form of systematic reviews and meta-analyses, generally considered reliable depending on their methodology and the reliability of the studies they include. Typically, the conservation literature is thought to have relatively few studies with reliable study designs, due to logistical, funding, and time-based constraints (Christie et al., 2019; de Palma et al., 2018). However, we do not know how this broad pattern varies geographically (i.e., are reliable study designs used more often in certain regions?), or the prevalence of these study designs in the literature that tests conservation interventions. To our knowledge, previous research (Burivalova et al., 2019) has only quantified this in the tropics for evidence on the effectiveness of tropical forest conservation strategies, and not on a global scale for a variety of conservation interventions. Insufficient reliable evidence in certain regions would mandate greater efforts to improve, where possible, the types of study design implemented in those locations.

The aim of this study is to improve our empirical and quantitative understanding of the biases and gaps in the evidence base for conservation. To do this, we present a series of analyses of the Conservation Evidence database (Sutherland et al., 2019), a comprehensive collection of 5,816 publications (as of March 2020) that have quantitatively tested the effectiveness of conservation interventions. To quantify bias in this evidence base we set out to answer several research questions for two taxa (amphibians and birds): 1) what is the geographic distribution

of studies?; 2) how does this distribution vary for studies with different designs?; 3) what is the taxonomic distribution of studies?; and for studies on a given conservation intervention, how much variation is there in the use of 4) different study designs, and 5) different metrics? Identifying patterns, biases, and knowledge gaps in the evidence base can help in setting priorities for future research. With a more reliable and complete evidence base, research can better support evidence-based decision-making in conservation and ultimately more effective conservation.

Materials and methods

Conservation Evidence database

The Conservation Evidence project summarises studies that have quantitatively tested the effect of a conservation intervention (Sutherland et al., 2019). Conservation interventions are defined as “actions that have been or could be used to conserve biodiversity”, and the effect that is quantified can be “on any aspect of biodiversity (e.g., abundance of a focal species, survival rates of translocated individuals, use of nest boxes, extent of habitat) or human behaviour related to biodiversity conservation (e.g., levels of hunting, or sales of products detrimental to biodiversity).” (Sutherland et al., 2019, p.3). These studies are found using systematic manual searches of the conservation literature, including over 290 English and 150 non-English language journals (Sutherland et al., 2019). The Conservation Evidence website (www.conservationevidence.com), as of March 2020, is structured into 2,105 different interventions (e.g., control invasive mammals on islands) contained within 16 synopses (e.g., Bird Conservation) and displays a summary of each study included, or multiple summaries if a study’s results apply to several interventions (e.g., both pond creation and translocation of amphibians). A list of interventions is created for each synopsis through consulting initial literature scans (but before systematic manual searches) and an advisory board (a range of academics, practitioners, and policymakers with subject-specific expertise from different parts of the world; Sutherland et al., 2019). Interventions are usually described at a fine scale (for example, “set longlines at the side of the boat to reduce seabird bycatch” is a separate intervention to “set lines underwater to reduce seabird bycatch”).

As we wanted to assess the number of studies per intervention for certain subsets of studies (e.g., by the metric or study design used), we grouped similar interventions that focused on single taxa or habitats (e.g., “create ponds for frogs” and “create ponds for toads” would be grouped into “create ponds”; see Table S1). This ensured that the scope of interventions was appropriate for our analysis and did not act as a constraint on the numbers of studies per intervention.

We extracted metadata from the database for every study within the amphibian (n=410; Smith and Sutherland, 2014) and bird synopses (n=1,239; Williams et al., 2013), including latitude and longitude coordinates (mean coordinates where a study used multiple sites). We only considered studies for amphibians and birds as these taxa had the most complete and comprehensive metadata in the database. The literature searches that retrieved these studies (Sutherland et al., 2019) were last conducted in 2012 for amphibians and 2011 for birds. Whilst these searches are not as recent as we might wish, these data provide the only way to reasonably assess biases using a large number of studies that have tested the effectiveness

of conservation interventions. For all analyses we excluded interventions that did not contain any studies (i.e., no studies present regardless of biome, metric, or design types; 31 interventions for Amphibians and 56 for birds).

Patterns in evidence for different metrics and designs

A standardised set of keywords are used to describe study design in the Conservation Evidence database (Table 1). A single report or paper summarised in the database may use multiple study designs if several tests are described. Each study design used within a report or paper constitutes an individual study, each of which were counted separately. An individual study can also be assigned to multiple interventions and multiple synopses if it contains relevant information. We used the number of studies per intervention as the major variable of interest. To determine the accuracy of reported study designs, we manually checked the original papers of a random 5% of studies in the database (n=21 for Amphibians; n=62 for Birds). The correct design was reported to 95% of amphibian studies (one study with an After design was misreported as a Before-After design; Table S2) and 94% of bird studies (one CI study misreported as After, one BACI study misreported as CI, two RCI studies misreported as CI; Table S2). As we were estimating the mean number of studies per intervention that used different study designs across many interventions, and the global geographical distribution of many studies using different designs (see next section), these misclassifications will have made little difference to these overall results.

Table 1 *– Definitions for each study design based on the criteria used to define them, and the keywords used, in the Conservation Evidence database (Sutherland et al. 2019). Experimental designs use randomised allocation of independent experimental units to treatment and control groups (RCI); quasi-experimental designs are not randomised but have a control group (Control-Impact or Before-After Control-Impact); and non-experimental designs lack a control group (Before-After or After).*

Criterion	**Design**				
	After	**Before-After**	**Control-Impact**	**Before-After Control-Impact**	**Randomised Control-Impact**
Design acronym		BA	CI	BACI	RCI
Control?	No	No	Yes	Yes	Yes
Sampling before intervention?	No	Yes	No	Yes	Yes or No
Randomised?	No	No	No	No	Yes
Matching or pairing?	No	No	Yes or No	Yes or No	-
Experimental?	Non-experimental (no comparison)	Non-experimental (no comparison)	Quasi-experimental	Quasi-experimental	Experimental
	Example				
	Monitoring the number of songbirds feeding in field margins after sowing wildflower seeds	Quantifying amphibian mortality on roads before and after creating road tunnels	Investigating invertebrate diversity in grazed and ungrazed grassland plots	Comparison between biodiversity before and after the addition of dead wood in streams using an upstream control and downstream treatment	Monitoring effect of crop rotation between randomly assigned treatment and control fields

To identify the metrics used by each study to measure the effectiveness of interventions, we first used web scraping to obtain summaries of studies from the Conservation Evidence website – using the XML package (Lang, 2019) and RCurl package (Lang and CRAN team, 2018) in R statistical software version 3.5.1 (R Core Team, 2019). We also used the doParallel package (Microsoft Corporation and Weston, 2019) to increase computational performance. Once summaries were obtained, we created and tested a set of regular expression rules (e.g., matching keywords and patterns; Appendix S1) to detect the following metric groups used by each study: 1. abundance, density, and cover; 2. mortality and survival; 3. diversity and species richness; 4. reproductive success. This was necessary as this information is currently unavailable in the database and allowed us to quantify the number of studies using each metric, and the number of unique metrics used, in each intervention.

For a random 5% of studies (n=21 amphibians, n=62 birds) we found that the metric groups identified by regular expressions were correct for 90% of amphibian studies and 95% of bird studies (Table S3). For amphibians, all misclassifications were false negatives (failure to detect abundance, density, and cover in two studies). For birds, there were false positives for two studies (3.2% – one erroneous detection of reproductive success and one of mortality/survival) and a false negative for one study (1.6% – failure to detect diversity and species richness). As we were using this automated classification to gain an overall estimate of the mean number of studies per intervention across a large number of interventions for each metric group, these misclassifications will have made little difference to these overall estimates. Automating the extraction of effectiveness metrics also offers the most feasible and reproducible methodology to analyse the entire evidence base and controls for some potential biases that would affect manual classification.

Patterns in evidence spatially and taxonomically

We mapped the spatial distribution of studies in the database by creating a raster layer with the raster package (Hijmans, 2020), summing the number of studies using different study designs for each 4x4-degree cell using longitude and latitude coordinates – we chose a 4x4-degree resolution to aid data visualisation for the maps we produced (Figs.1 & 2). We excluded reviews from our analyses as they were often global or regional in scale. To estimate the geographical coverage of studies we counted the number of countries and continents they were present in. We also compared the number of studies in each 2x2-degree cell with the number of species, threatened species and data-deficient species for extant amphibian and bird species using data downloaded from the International Union for Conservation of Nature (IUCN) Red List (IUCN, 2019). We chose a 2x2-degree grid cell resolution as this was the maximum appropriate resolution recommended by Hurlbert and Jetz (2007) when using range map data. We excluded grid cells containing zero studies and zero species and normalised the number of studies and species to between 0-1: $studies = (studies - studies_{min}) / (studies_{max} - studies_{min})$. We then quantified the relationship between the normalised number of studies (as the response variable) and species (as the explanatory variable) in each grid cell using a generalised linear model with a binomial error distribution and log link function. We repeated this normalisation and modelling separately for the number of threatened species and the number of data deficient species. A square root transformation of the explanatory variable (number of species, threatened species or data deficient species) did not substantially improve model fit (AIC values were not reduced by more than two units and R^2 values remained unchanged or only marginally increased; Table S4). We therefore chose untransformed models as these were more parsimonious. All modelling assumptions held in

terms of no overdispersion, and no substantial patterns between residuals and the explanatory variable or fitted values.

We assessed the relative under or overrepresentation of different biomes in the database by calculating the difference between the percentage of studies conducted in each biome and the percentage of the Earth's terrestrial area covered by each biome (Dinerstein et al., 2017). We assigned studies to each biome using longitude and latitude coordinates for each study, a shapefile of 14 terrestrial biomes (Dinerstein et al., 2017), and the sp package in R (Bivand et al., 2013; Pebesma and Bivand, 2005). We excluded studies conducted outside terrestrial biomes (e.g., studies considering seabirds over oceans).

To investigate the distribution of evidence taxonomically, we found the percentage of studies that tested an intervention on each of the major bird orders based on a cladogram from Prum et al. (2015). For amphibians, we did the same for the three major amphibian orders using a trimmed cladogram from Pyron and Wiens (2011). To investigate the representation of taxonomic orders in the evidence base, we calculated the difference between the proportion of studies: and 1) the proportion of threatened species in each order (relative to the number of all threatened amphibian or bird species); and 2) the proportion of amphibian and bird species in each order (relative to the number of all amphibian or bird species). We obtained data on the number of species and threatened species (with vulnerable, endangered, or critically endangered status) in each order from the IUCN Red List (IUCN, 2019).

All data analysed in this study and code to repeat analyses are available from www.doi.org/10.5281/zenodo.3634779.

Results

There was substantial bias in the spatial distribution of evidence on conservation interventions. Approximately 90% of amphibian studies and 84% of bird studies were conducted in either North America, Europe, or Australasia. Additionally, 64% of amphibian studies and 63% of bird studies were conducted in just three countries: the United Kingdom, United States and Australia. There were large spatial gaps in evidence in South America, Africa, Asia, and Russia for both amphibians and birds. There were also few studies in the tropics or close to the poles (Figs.1 & 2).

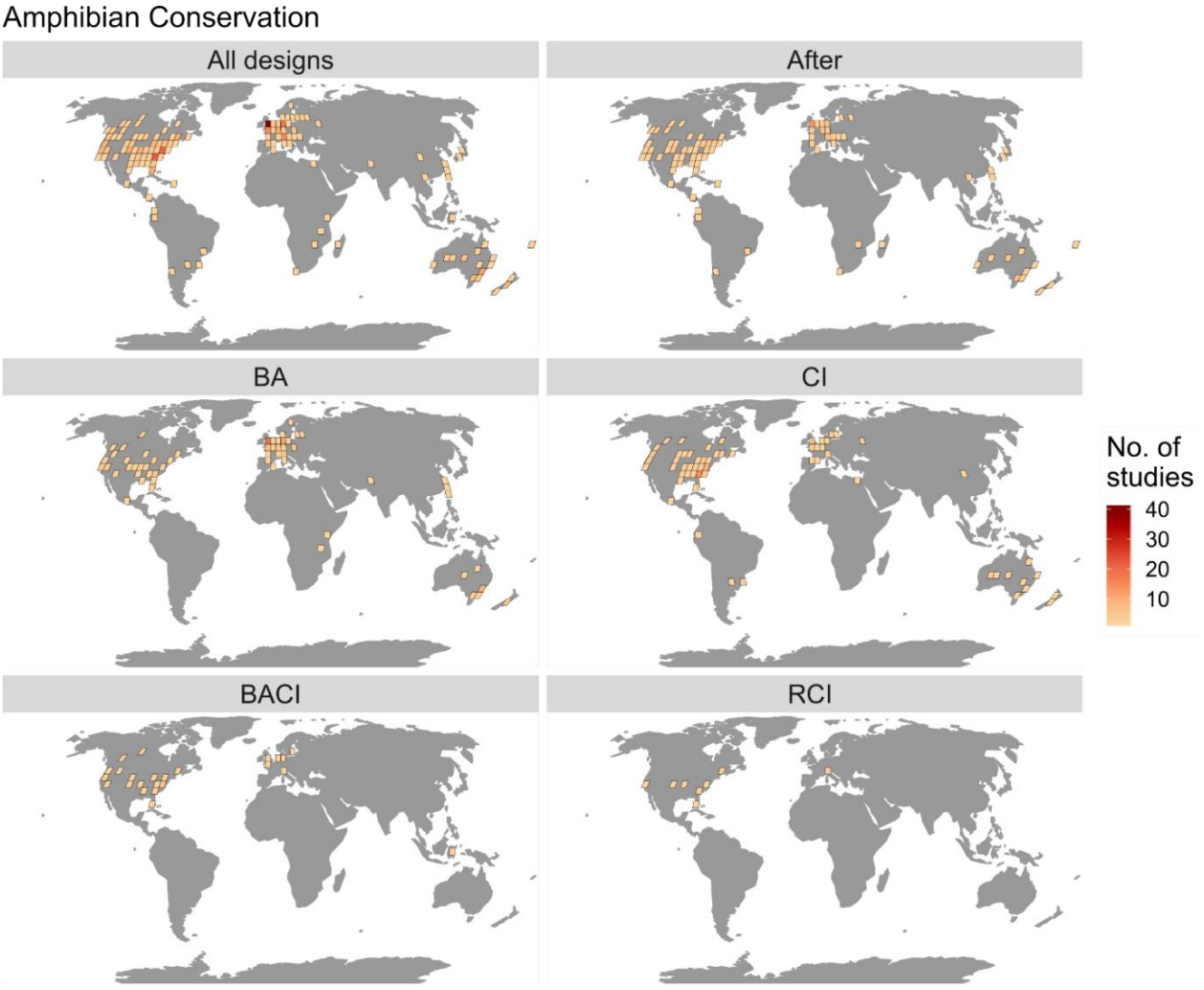


Figure 1 *– Spatial distribution of studies for amphibians using a Robinson projection and grid cells at a 4x4-degree resolution. Definitions of design acronyms are as follows: BA = Before-After; CI = Control-Impact; BACI = Before-After Control-Impact; RCI = Randomised Control-Impact (see Table 1 for details of designs).*

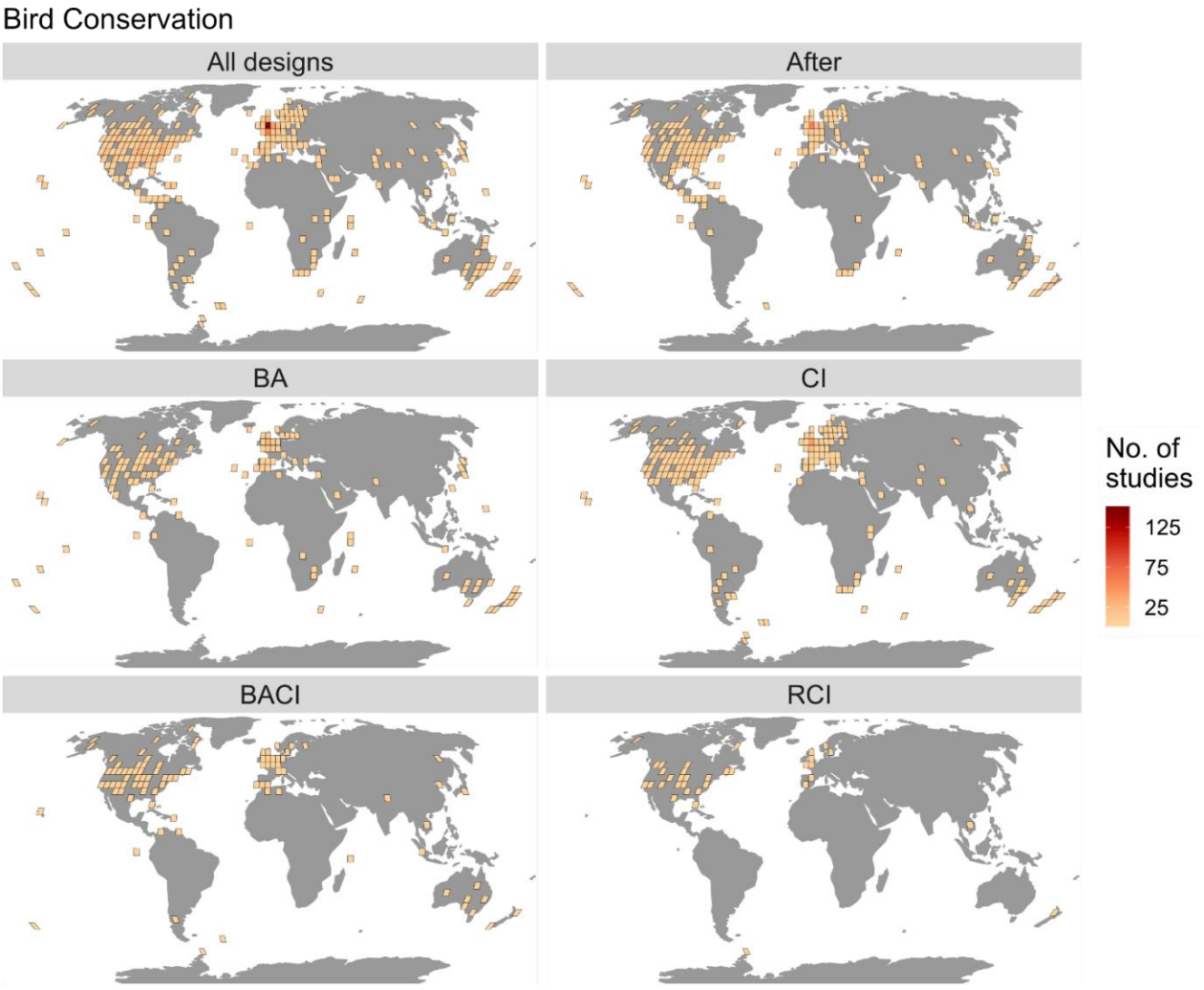


Figure 2 *– Spatial distribution of studies for birds using a Robinson projection and grid cells at a 4x4-degree resolution. Definitions of design acronyms are as follows: BA = Before-After; CI = Control-Impact; BACI = Before-After Control-Impact; RCI = Randomised Control-Impact (see Table 1 for details of designs).*

The geographical distribution of studies varied considerably by study design. Amphibian studies with the most reliable study designs, BACI and RCI, were concentrated in North America and Europe; these designs were almost absent from the tropics (Fig.1). No BACI or RCI studies for amphibians were conducted in South America, Africa, or Australasia (as well as Asia for RCI studies), and both types of study design were only used in six countries (Table S5; Fig.1). BA studies for amphibians were found in 23 countries (but none from South America), whilst CI studies were found in fewer countries (18) but were present in all continents where amphibians exist. Amphibian studies using After designs covered the greatest number of countries (31) across all possible continents (Table S5).

The evidence for birds had a greater geographical coverage than for amphibians, particularly in the tropics (Fig.2). No RCI studies were conducted in Antarctica or South America, whilst studies using other designs were present in all seven continents. RCI and BACI studies were also present in considerably fewer countries than After, CI, and BA studies (Table S5).

There was no statistically significant spatial relationship ($p=0.37$; Table S6) between the number of studies and the number of amphibian species, and a small, but statistically significant positive spatial relationship with the number of bird species ($p<0.01$; Fig.3; Table S6). Conversely, the number of studies significantly decreased with the number of threatened species (birds: $p<0.01$; amphibians: $p=0.03$) and data deficient species (birds: $p=0.03$; amphibians: $p=0.04$) – however, the magnitude of this decrease was small for birds (Fig.3; Table S6). For amphibians, the grid cell with the most studies (normalised value of 1; Fig.3) covered central England, whilst for birds, the two grid cells with the most studies covered central and northern England (normalised values of 0.95 and 1, respectively; Fig.3).

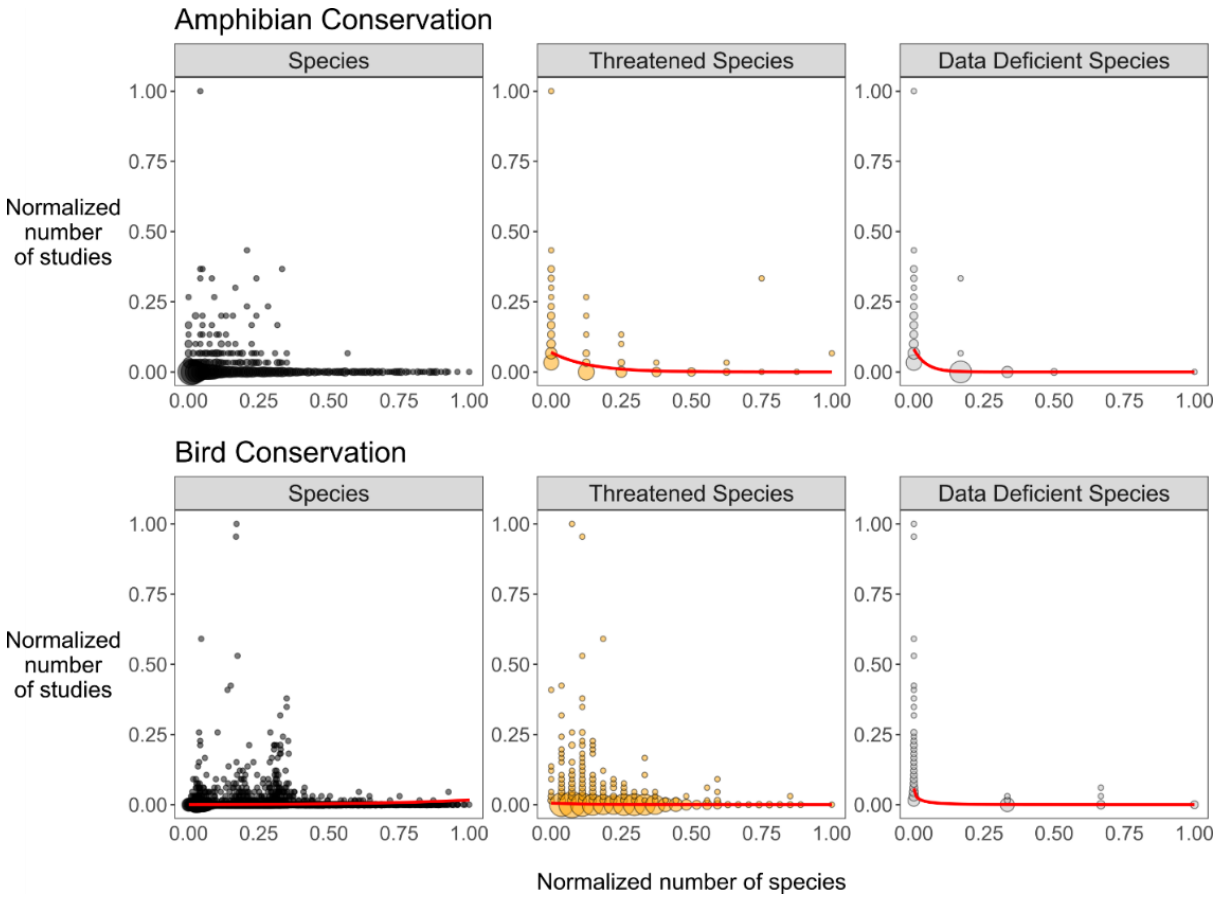


Figure 3 *– Comparison of the normalised number of studies and the number of species (all species, threatened species and data deficient species) in 2x2-degree grid cells for amphibians and birds (with one representing the cells with the most studies or species and zero being the fewest). Cells with both zero studies and zero species were excluded. Red fitted lines are based on binomial generalised linear models where statistically significant increases or decreases were detected (p<0.05; see Materials and methods for details). Note that the slopes of the regression lines were negative for threatened and data deficient amphibian and bird species. The size of points is proportional to the number of points at that position on the figure to aid visualisation. Threatened species are those classified as vulnerable, endangered, or critically endangered respectively, according to the IUCN Red List (IUCN 2019).*

There was also substantial variation in the representation of different amphibian and bird orders in the evidence base relative to the proportion of threatened species each order contained. For birds, the most well represented orders were shorebirds (Charadriiformes), waterfowl (Anseriformes), and falcons (Falconiformes), respectively – i.e., high proportions of studies relative to proportions of threatened species (Fig.4). Songbirds (Passeriformes), parrots (Psittaciformes), pigeons (Columbiformes), and nightjars, hummingbirds, and swifts (Caprimulgiformes), were the least well represented bird orders – i.e., low proportions of studies relative to threatened species. No studies were present for several bird orders, such as hornbills and hoopoes (Bucerotiformes; see names in red in Figure 4). For amphibians, frogs (Anura) were the least well represented, whilst salamanders (Caudata) were the most well represented. There was only a single study for the entire order of Caecilians (Gymnophiona; Fig.4). Patterns were different when considering the proportion of studies relative to the proportion of species in each bird order; most bird orders were relatively well represented apart from songbirds and orders for which there were no studies (Figure S1). For amphibians, patterns in representation were similar for both the proportion of species and proportion of threatened species (Figure S1; Figure 4).

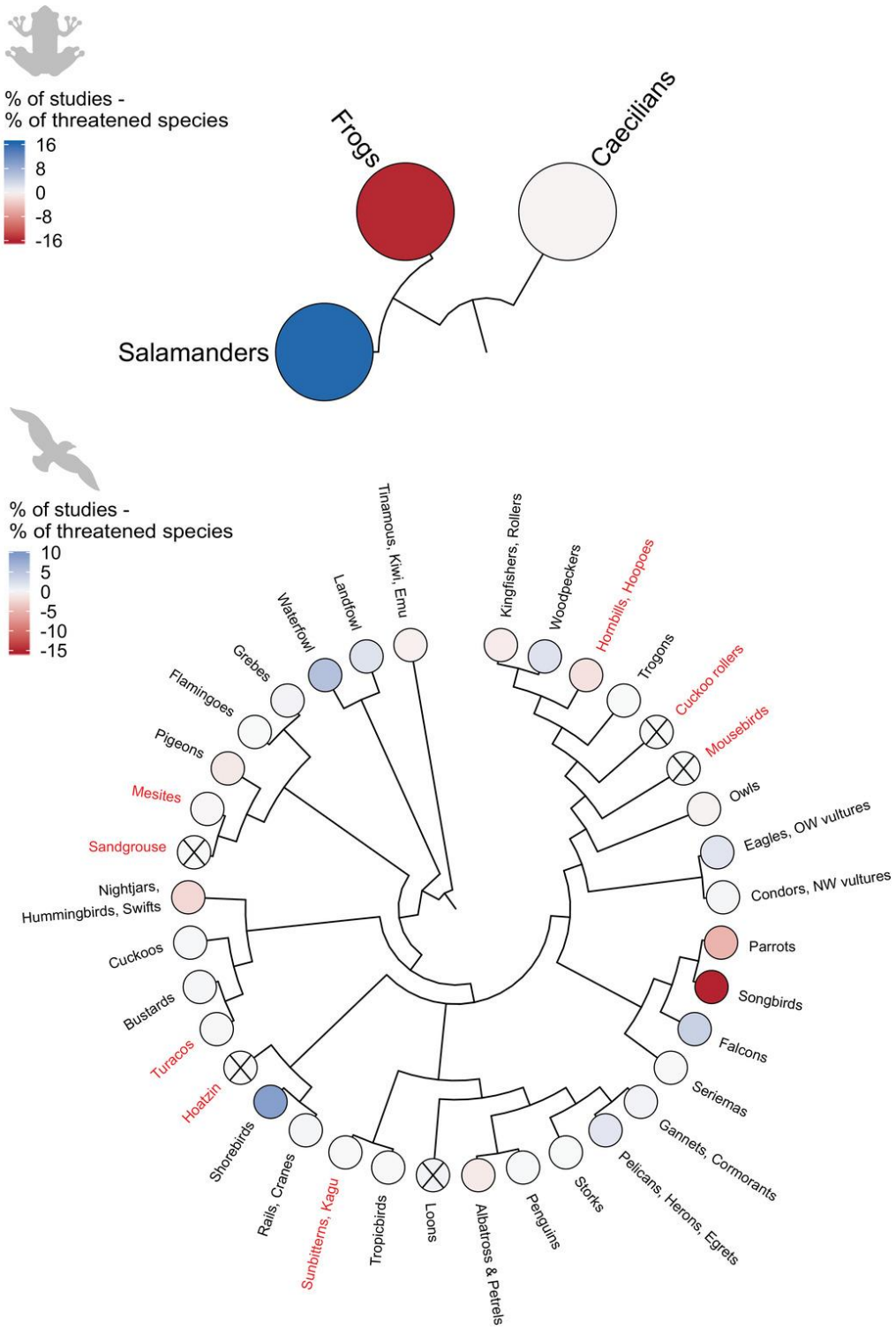


Figure 4 *– Percentage of studies minus percentage of threatened species in each order of amphibians and birds – percentages are relative to the total number of amphibian or bird studies and species. Red names indicate zero studies for that order and black crosses indicate that order contains zero threatened species. Darker blue colours represent higher proportions of studies relative to the proportion of threatened species, whilst darker red colours indicate relatively lower proportions of studies.*

Certain biomes were better represented (in terms of the total number of studies conducted in each biome) relative to the percentage of the Earth's terrestrial area they covered – notably Temperate Broadleaf & Mixed Forests, Temperate Grasslands, Savannas & Shrublands, Temperate Conifer Forests, and Mediterranean Forests, Woodlands & Scrub for both amphibians and birds (Fig.5). The three most underrepresented biomes for both amphibians and birds were Deserts & Xeric Shrublands, Tropical & Subtropical Grasslands, Savannas & Shrublands, and Tropical & Subtropical Moist Broadleaf Forests (Fig.5). For amphibians, there were no studies in Tropical & Subtropical Coniferous Forests, Tropical & Subtropical Dry Broadleaf Forests, and Tundra (red outlines to circles in Figure 5).

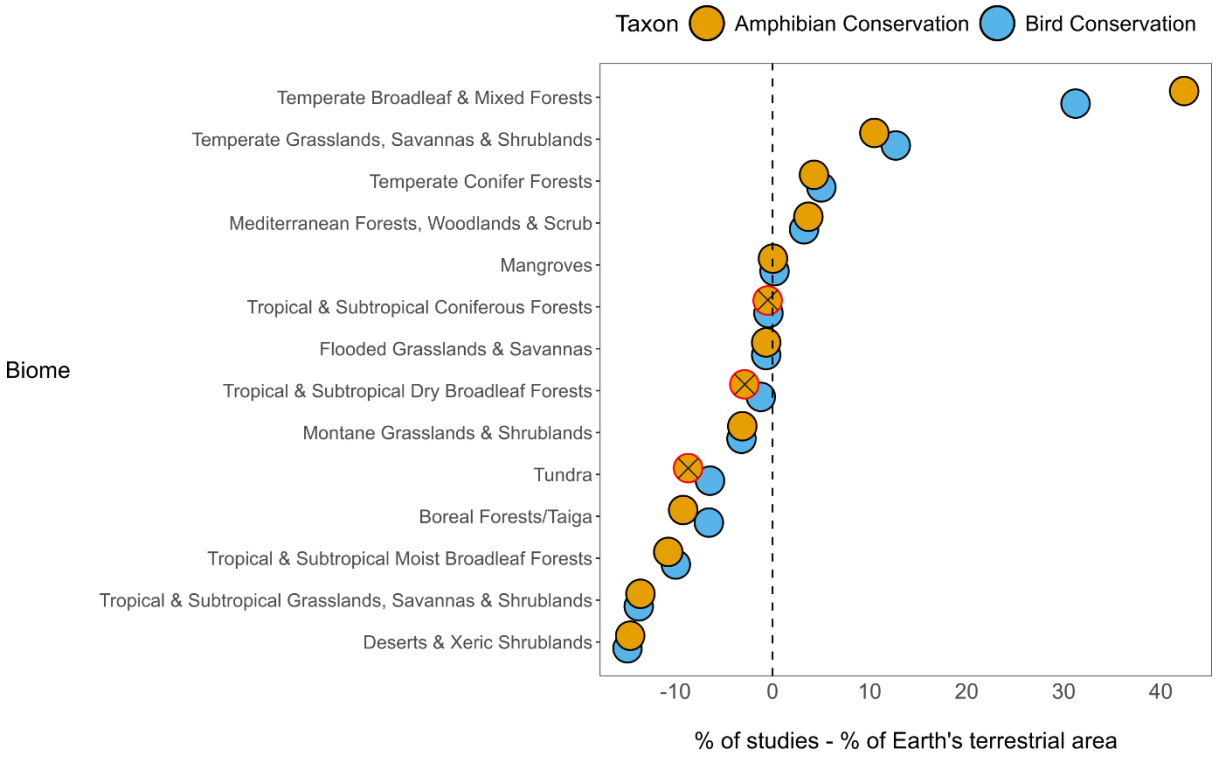


Figure 5 *– Percentage of amphibian and bird studies conducted in each biome minus the percentage of Earth's terrestrial area covered by each biome. Circles with red outlines and crosses through them indicate that no studies were conducted in that biome.*

The total number of interventions (containing at least one study) was 243 for birds and 74 for amphibians. On average, there were more studies per intervention for amphibians than for birds (although the total number of studies was greater for birds than amphibians). There was a higher proportion of interventions for birds that contained one study (34%) than amphibians (24%) – i.e., a more right skewed distribution of studies per intervention for birds than amphibians (Figure S2).

The most commonly used metrics in amphibian conservation were mortality/survival (3.9 studies per intervention) and reproductive success (3.8 studies per intervention), whilst for

birds, mortality/survival (3.9 studies per intervention) and abundance/density/cover (3.8 studies per intervention; Figure S3) were the most common. On average, the effectiveness of each intervention was measured using 2.1 different metrics for amphibians and 3.3 metrics for birds.

There was a low number of studies per intervention that used reliable BACI or RCI designs (fewer than 0.3 studies per intervention for both amphibians and birds; Figure S3). Studies most commonly used the least reliable After design, followed by CI and BA designs, for both amphibians and birds – note declines in number of studies per intervention when excluding studies using certain designs; Figure S3).

Discussion

Our work demonstrates that the evidence base for amphibian and bird conservation suffers from severe geographical and taxonomic biases that may hamper our ability to make locally relevant evidence-based recommendations to decision-makers. Geographically, studies were concentrated in North America, Europe, and Australasia, and there were negative spatial relationships between the number of studies and the number of threatened species and data deficient species for both taxa. That the most well represented biomes in the evidence base were Temperate Broadleaf & Mixed Forests and Temperate Grasslands, Savannas & Shrublands also indicated strong geographic bias. Taxonomically, certain orders were better studied relative to the number of threatened species they contained (e.g., salamanders for amphibians and shorebirds, falcons, and waterfowl for birds), whilst some orders were not studied at all (e.g., hornbills and hoopoes).

These results show even more severe geographic biases than other studies of the wider conservation literature; the clear paucity of evidence from the polar regions (expected for amphibians but concerning for birds), Africa, Russia, the Middle East, and South America appear more severe than those shown by di Marco et al. (2017), Hickisch et al. (2019), and Wilson et al. (2016). The United Kingdom also rivalled the United States as a hotspot of evidence for these two taxonomic groups, which was not as apparent in Wilson et al. (2016) or Hickisch et al. (2019) but was in di Marco et al. (2017). This hotspot contrasts, particularly for amphibians, with their low species richness in the United Kingdom (only seven native amphibian species). In their review of the effectiveness of terrestrial protected areas, Geldmann et al. (2013) found different geographic biases, away from North America and Europe towards Latin America, Africa, and Asia. We believe this is because we considered a different subset of studies, focusing only on studies that had quantitatively tested a variety of conservation interventions, as opposed to the effectiveness of terrestrial protected areas.

That the number of studies testing conservation interventions had a negative relationship with the number of threatened species and data deficient species is also concerning. This pattern has not previously been found by studies of the wider conservation literature, which instead reported positive relationships with the number of threatened species in the tropics (Reboredo Segovia et al., 2020). Such patterns clearly suggest that greater research effort needs to be targeted at testing conservation actions in regions with large numbers of threatened species that urgently require effective conservation.

However, we must acknowledge that some of the geographic bias we found could be attributable to the low number of studies from non-English language journals that are currently included in the Conservation Evidence database. Publications from over 317 journals

published in 10 languages are being added to the database through the Transcending Language Barriers to Environmental Sciences project (translatE). However, language bias is a common problem affecting most scientific evidence syntheses (Neimann Rasmussen and Montgomery, 2018) that is often ignored. As researchers conducting evidence synthesis, we must do more to seek out and collate evidence published in non-English languages and the grey literature. This is particularly important given that approximately 36% of the wider conservation literature is found in non-English language journals (Amano et al., 2016). However, where non-English literature was included in Conservation Evidence searches (e.g., relevant ecology and conservation journals in Portuguese and Spanish for the Bat Conservation synopsis), the percentage of studies testing conservation actions was small at 0.4% (six studies out of 1,492 studies systematically searched; Berthinussen et al., 2014). More generally, for all non-English journals searched to date for Conservation Evidence (across all synopses), the verified rate of studies testing conservation actions is smaller at 0.18% or 643 studies out of 345,119 (S. Petrovan, personal communication, March 20 2020). This suggests that few studies testing conservation actions would be added from the non-English literature – possibly because a substantial proportion of non-English studies may describe conservation threats and ecology, rather than quantitatively test conservation actions. Therefore, language bias is unlikely to have substantially affected the broad patterns in our results. However, non-English studies that test conservation actions are potentially the only available studies for certain species and geographical areas (Berthinussen et al., 2014) and so it is still important to synthesise these studies to inform future conservation efforts.

Some taxonomic orders were well represented in the evidence on conservation effectiveness relative to the percentage of threatened species in each order, whilst other orders were very poorly represented (Fig.4) – as found in analyses of the wider conservation literature (Clark and May, 2002; Donaldson et al., 2016; Fazey et al., 2005; Murray et al., 2015). Most bird species and thus most threatened bird species were songbirds (Passeriformes; 46% of all threatened bird species) but this order was represented the poorest (31% of studies), followed by parrots (Psittaciformes; 8% of all threatened bird species but only 2% of studies). Conversely, shorebirds (Charadriiformes) and waterfowl (Anseriformes) were the most well represented (3% and 2% of all threatened bird species and 13% and 8% of studies, respectively). These differences in representation probably reflect the relative difficulty in studying threatened songbird species (e.g., small-bodied, forest species with small range sizes) and parrots (often found in less easily accessible tropical locations), compared to shorebirds and waterfowl (which generally have larger range sizes that often overlap with hotspots of research effort in North America and Europe).

Among amphibians, salamanders were well represented as they contained only 14% of all threatened amphibian species, and yet appeared in 30% of studies. This is potentially because certain non-threatened, protected species such as Great Crested Newts (*Triturus cristatus*) (a European protected species with an IUCN Red List status of Least Concern) are highly studied in relation to the effectiveness of mitigation interventions and that one third of salamander species exist in North America where research effort is concentrated. Frogs (Anura) were underrepresented (70% of studies versus ~86% of threatened amphibian species) possibly because many threatened frog species exist in less easily accessible tropical locations. Caecilians (Gymnophiona) were only represented by a single study, but this was in proportion to the number of threatened species they represent (only ~0.6% of all threatened amphibian species).

An underrepresentation of threatened species is concerning because information on the effectiveness of interventions targeting threatened species is urgently required – particularly given substantial declines of bird fauna (Rosenberg et al., 2019) and severe threats to amphibians (Grant et al., 2019). Whilst it can be challenging to design reliable studies on rare species, where feasible, conservation scientists should prioritise testing the effectiveness of conservation interventions for threatened species. Equally, the absence of some orders from the literature testing conservation interventions is problematic, because functional and ecological differences between taxonomic groups may make generalisation of the effectiveness of interventions difficult or inappropriate. Investigating which interventions are likely to be effective in many local contexts is extremely important to prioritise the most important taxonomic gaps in evidence that need to be addressed.

Types of bias that may complicate the process of evidence synthesis were also present; for example, studies with more reliable designs (e.g., RCI or BACI) tended to be strongly concentrated in North America and Europe (particularly the United Kingdom) compared to less reliable designs (e.g., BA, CI, and After designs; Figs.1 & 2). Combined geographic and study design bias has not been previously found (e.g., Burivalova et al. (2019) did not find patterns across continents in the tropical forest conservation literature) and suggests that we not only lack studies outside of North America and Europe, but that the few studies that do exist outside these regions are likely to be of low reliability (Christie et al., 2019). This may be because studies conducted outside North America and Europe face greater constraints – e.g., logistical, funding, and time constraints – on the types of study design they can use when assessing the effectiveness of conservation actions. Therefore, funders, journals, and researchers need to facilitate tests of conservation interventions using reliable study designs in these underrepresented regions and the publication of their results.

Amphibian and bird studies also used a variety of metrics to quantify the effectiveness of the same intervention. Although using several metrics may help us better understand the overall effectiveness of an intervention, too many could make evidence hard to synthesise in systematic reviews and meta-analyses (by reducing the number of directly comparable studies), and difficult to interpret for decision-makers. This highlights the need for greater standardisation of the sets of metrics used to assess the effectiveness of certain interventions (Capmourteres and Anand, 2016; McQuatters-Gollop et al., 2019) to help make studies more directly comparable.

The gaps and biases we have highlighted in the literature on the effectiveness of conservation interventions represents a serious issue for the field of conservation. Although we could only analyse the literature up until 2012 for amphibians and 2011 for birds, these gaps and biases are still likely to persist. However, with limited resources we cannot afford to allocate research effort inefficiently. Our results are therefore extremely important for determining where future research effort on testing the effectiveness of conservation interventions should be invested. Future studies should not only focus on testing conservation interventions on the poorly represented threatened taxa, regions, and biomes we identified, using reliable study designs where possible, but also other poorly represented taxa that Conservation Evidence is beginning to, or has yet to, summarise the evidence on (e.g., plants, insects, and reptiles). Future work could also identify whether there are system- or species-specific interventions that are not included in the Conservation Evidence database, particularly in relatively poorly studied regions. Interventions are defined by an advisory board before systematic literature searches occur but are often updated and reframed when studies are found to mention or test additional interventions – listed interventions therefore reflect those described in the conservation literature. Whilst possible bias in interventions does not affect the inclusion of studies in the database (as studies are included in the database if they quantitatively test any conservation intervention), identifying possible interventions that are not listed at www.conservationevidence.com would be useful to prioritise the testing of future interventions, particularly for underrepresented regions or taxa. This work would also benefit from a more systematic, hierarchical classification system for describing interventions.

Future work is needed to identify specific research priorities for testing conservation interventions for taxonomic groups other than amphibians and birds, although the broad biases we identified here are likely to apply to other taxa. We hope that by addressing geographic and taxonomic biases in the evidence base for conservation we can ensure more relevant evidence-based recommendations can be made to decision-makers. Similarly, addressing the geographic bias in the use of reliable study designs, and in the variability in the types of metrics used by studies, will hopefully allow evidence synthesis to be more

efficient and effective. A more complete, reliable, and standardised evidence base will enable conservation to become more evidence-based and, ultimately, more effective.

References

Amano, T., González-Varo, J.P., Sutherland, W.J., 2016. Languages Are Still a Major Barrier to Global Science. PLOS Biology 14, e2000933. https://doi.org/10.1371/journal.pbio.2000933

Amano, T., Sutherland, W.J., 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. Proceedings of the Royal Society B: Biological Sciences 280, 20122649. https://doi.org/10.1098/rspb.2012.2649

Aranda, S.C., Gabriel, R., Borges, P.A. v, Azevedo, E.B. de, Lobo, J.M., 2011. Designing a survey protocol to overcome the Wallacean shortfall: a working guide using bryophyte distribution data on Terceira Island (Azores). The Bryologist 114, 611–624. https://doi.org/10.1639/0007-2745-114.3.611

Berthinussen, A., Richardson, O.C., Altringham, J.D., 2014. Bat conservation: global evidence for the effects of interventions. Pelagic Publishing Ltd., Exeter.

Bivand, R.S., Pebesma, E., Gomez-Rubio, V., 2013. Applied spatial data analysis with R, Second ed. Springer, New York.

Brodie, J.F., 2009. Is research effort allocated efficiently for conservation? Felidae as a global case study. Biodiversity and Conservation 18, 2927–2939. https://doi.org/10.1007/s10531-009-9617-3

Brooke, Z.M., Bielby, J., Nambiar, K., Carbone, C., 2014. Correlates of Research Effort in Carnivores: Body Size, Range Size and Diet Matter. PLOS ONE 9, e93195. https://doi.org/10.1371/journal.pone.0093195

Burivalova, Z., Miteva, D., Salafsky, N., Butler, R.A., Wilcove, D.S., 2019. Evidence Types and Trends in Tropical Forest Conservation Literature. Trends in Ecology and Evolution 34, 669–679. https://doi.org/10.1016/j.tree.2019.03.002

Capmourteres, V., Anand, M., 2016. “Conservation value”: a review of the concept and its quantification. Ecosphere 7, e01476. https://doi.org/10.1002/ecs2.1476

Cardoso, P., Erwin, T.L., Borges, P.A. v, New, T.R., 2011. The seven impediments in invertebrate conservation and how to overcome them. Biological Conservation 144, 2647–2655. https://doi.org/https://doi.org/10.1016/j.biocon.2011.07.024

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Clark, J.A., May, R.M., 2002. Taxonomic Bias in Conservation Research. Science 297, 191–192. https://doi.org/10.1126/science.297.5579.191b

de Palma, A., Sanchez-Ortiz, K., Martin, P.A., Chadwick, A., Gilbert, G., Bates, A.E., Börger, L., Contu, S., Hill, S.L.L., Purvis, A., 2018. Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity: Study Design, Time, Space and Synthesis. Next Generation Biomonitoring 58, 163–199. https://doi.org/https://doi.org/10.1016/bs.aecr.2017.12.004

di Marco, M., Chapman, S., Althor, G., Kearney, S., Besancon, C., Butt, N., Maina, J., Possingham, H., Rogalla von Bieberstein, K., Venter, O., 2017. Changing trends and persisting biases in three decades of conservation science. Global Ecology and Conservation 10, 32–42. https://doi.org/10.1016/j.gecco.2017.01.008

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.P.B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F., Saleem, M., 2017. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. BioScience 67, 534–545. https://doi.org/10.1093/biosci/bix014

Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J.B., Collen, B., 2014. Defaunation in the Anthropocene. Science 345, 401–406. https://doi.org/10.1126/science.1251817

Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J., Kerr, J.T., 2016. Taxonomic bias and international biodiversity conservation research. FACETS 1, 105–113. https://doi.org/10.1139/facets-2016-0011

Fazey, I., Fischer, J., Lindenmayer, D.B., 2005. What do conservation biologists publish? Biological Conservation 124, 63–73. https://doi.org/https://doi.org/10.1016/j.biocon.2005.01.013

Geldmann, J., Barnes, M., Coad, L., Craigie, I.D., Hockings, M., Burgess, N.D., 2013. Effectiveness of terrestrial protected areas in reducing habitat loss and population declines. Biological Conservation 161, 230–238. https://doi.org/https://doi.org/10.1016/j.biocon.2013.02.018

Grant, E.H.C., Muths, E., Schmidt, B.R., Petrovan, S.O., 2019. Amphibian conservation in the Anthropocene. Biological Conservation 236, 543–547. https://doi.org/https://doi.org/10.1016/j.biocon.2019.03.003

Hickisch, R., Hodgetts, T., Johnson, P.J., Sillero-Zubiri, C., Tockner, K., Macdonald, D.W., 2019. Effects of publication bias on conservation planning. Conservation Biology 33, 1151–1163. https://doi.org/10.1111/cobi.13326

Hijmans, R.J., 2020. raster: Geographic Data Analysis and Modeling. R package version 3.3-13.

Hurlbert, A.H., Jetz, W., 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences of the United States of America 104, 13384–13389. https://doi.org/10.1073/pnas.0704469104

IUCN, 2019. IUCN Red List [WWW Document]. URL https://www.iucnredlist.org/ (accessed 11.12.19).

Lang, D.T., 2019. XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.99-0.5.

Lang, D.T., CRAN team, 2018. RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.95-4.11.

Martin, L.J., Blossey, B., Ellis, E., 2012. Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. Frontiers in Ecology and the Environment 10, 195–201. https://doi.org/https://doi.org/10.1890/110154

McQuatters-Gollop, A., Mitchell, I., Vina-Herbon, C., Bedford, J., Addison, P.F.E., Lynam, C.P., Geetha, P.N., Vermeulan, E.A., Smit, K., Bayley, D.T.I., Morris-Webb, E., Niner, H.J., Otto, S.A., 2019. From Science to Evidence – How Biodiversity Indicators Can Be Used for Effective Marine Conservation Policy and Management. Front. Mar. Sci. 6, 1–16. https://doi.org/10.3389/fmars.2019.00109

Meyer, C., Jetz, W., Guralnick, R.P., Fritz, S.A., Kreft, H., 2016. Range geometry and socio-economics dominate species-level biases in occurrence information. Global Ecology and Biogeography 25, 1181–1193. https://doi.org/https://doi.org/10.1111/geb.12483

Microsoft Corporation, Weston, S., 2019. doParallel: Foreach Parallel Adaptor for the “parallel” Package. R package version 1.0.15.

Murray, H.J., Green, E.J., Williams, D.R., Burfield, I.J., de Brooke, M.L., 2015. Is research effort associated with the conservation status of European bird species? Endangered Species Research 27, 193–206. https://doi.org/10.3354/esr00656

Neimann Rasmussen, L., Montgomery, P., 2018. The prevalence of and factors associated with inclusion of non-English language studies in Campbell systematic reviews: a survey and meta-epidemiological study. Systematic Reviews 7, 129. https://doi.org/10.1186/s13643-018-0786-6

Nuñez, M.A., Barlow, J., Cadotte, M., Lucas, K., Newton, E., Pettorelli, N., Stephens, P.A., 2019. Assessing the uneven global distribution of readership, submissions and publications in applied ecology: Obvious problems without obvious solutions. Journal of Applied Ecology 56, 4–9. https://doi.org/10.1111/1365-2664.13319

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569. https://doi.org/10.1038/nature15697

Pyron, R.A., Wiens, J.J., 2011. Molecular Phylogenetics and Evolution A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Molecular Phylogenetics and Evolution 61, 543–583. https://doi.org/10.1016/j.ympev.2011.06.012

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Reboredo Segovia, A.L., Romano, D., Armsworth, P.R., 2020. Who studies where? Boosting tropical conservation research where it is most needed. Frontiers in Ecology and the Environment fee.2146. https://doi.org/10.1002/fee.2146

Reddy, S., Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30, 1719–1727. https://doi.org/https://doi.org/10.1046/j.1365-2699.2003.00946.x

Rosenberg, K. v, Dokter, A.M., Blancher, P.J., Sauer, J.R., Smith, A.C., Smith, P.A., Stanton, J.C., Panjabi, A., Helft, L., Parr, M., Marra, P.P., 2019. Decline of the North American avifauna. Science eaaw1313. https://doi.org/10.1126/science.aaw1313

Smith, R.K., Sutherland, W.J., 2014. Amphibian conservation: global evidence for the effects of interventions. Pelagic Publishing Ltd., Exeter.

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19, 305–308. https://doi.org/https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Williams, D.R., Pople, R.G., Showler, D.A., Dicks, L. v, Child, M.F., Zu Ermgassen, E.K.H.J., Sutherland, W.J., 2013. Bird Conservation: Global evidence for the effects of interventions. Pelagic Publishing Ltd., Exeter.

Wilson, K.A., Auerbach, N.A., Sam, K., Magini, A.G., Moss, A.St.L., Langhans, S.D., Budiharta, S., Terzano, D., Meijaard, E., 2016. Conservation Research Is Not Happening Where It Is Most Needed. PLOS Biology 14, e1002413. https://doi.org/10.1371/journal.pbio.1002413

Supplementary Information

Figure S1

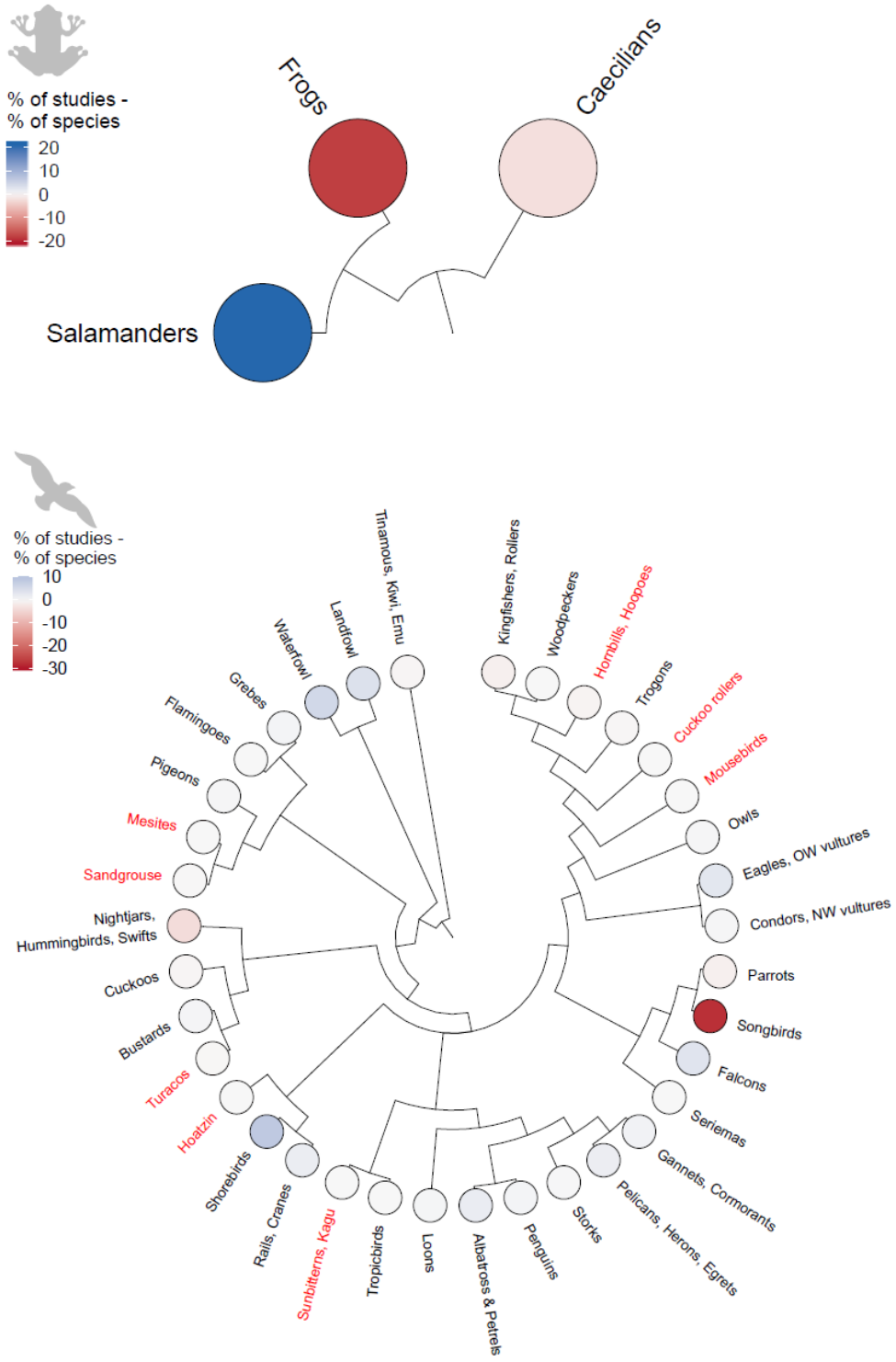


Figure S1 *– Percentage of studies minus percentage of species in each order of amphibians and birds (percentages relative to the total number of amphibian or bird studies and species) (red, 0 studies for that order; dark blue, high proportions of studies relative to the proportion of species; dark red, relatively lower proportions of studies).*

Figure S2

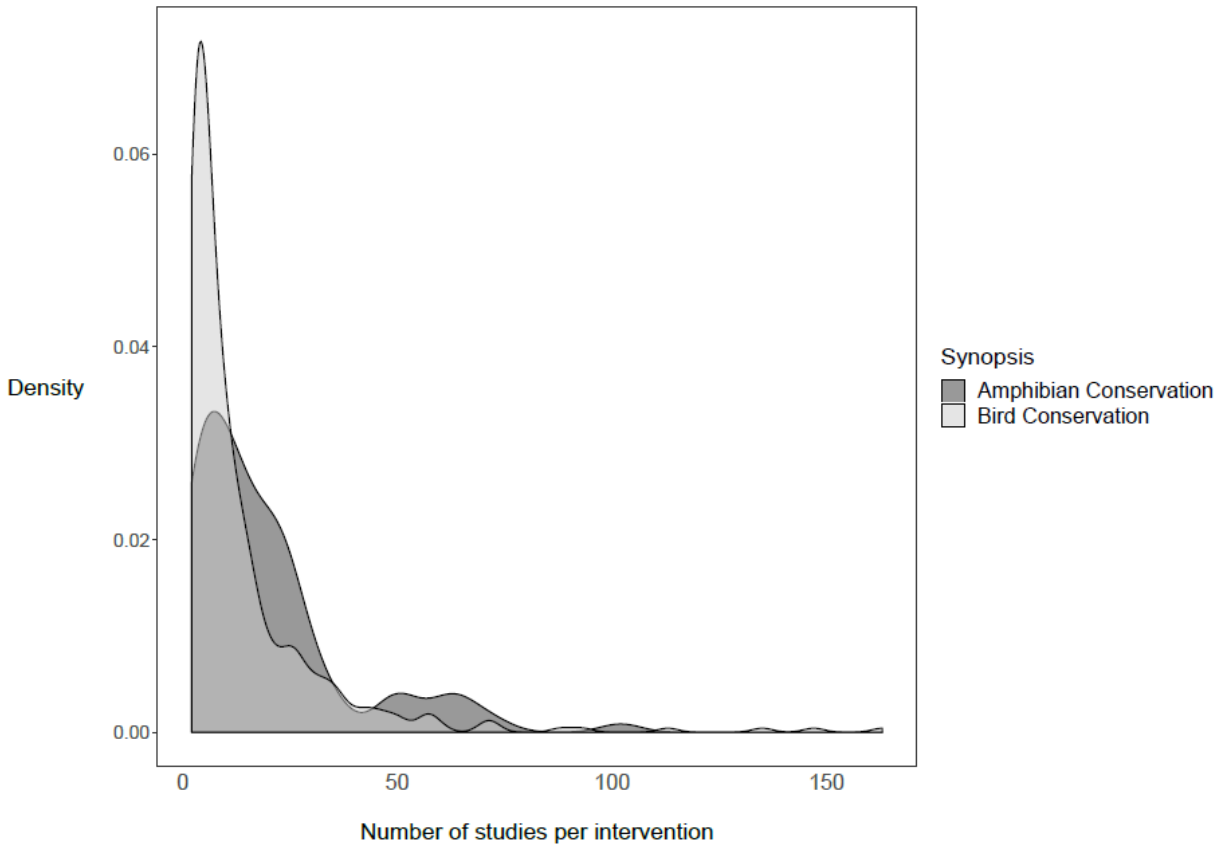


Figure S2 *– Distribution of the number of studies in each conservation intervention for amphibians and birds based on the Conservation Evidence database. Interventions containing zero studies were excluded.*

Figure S3

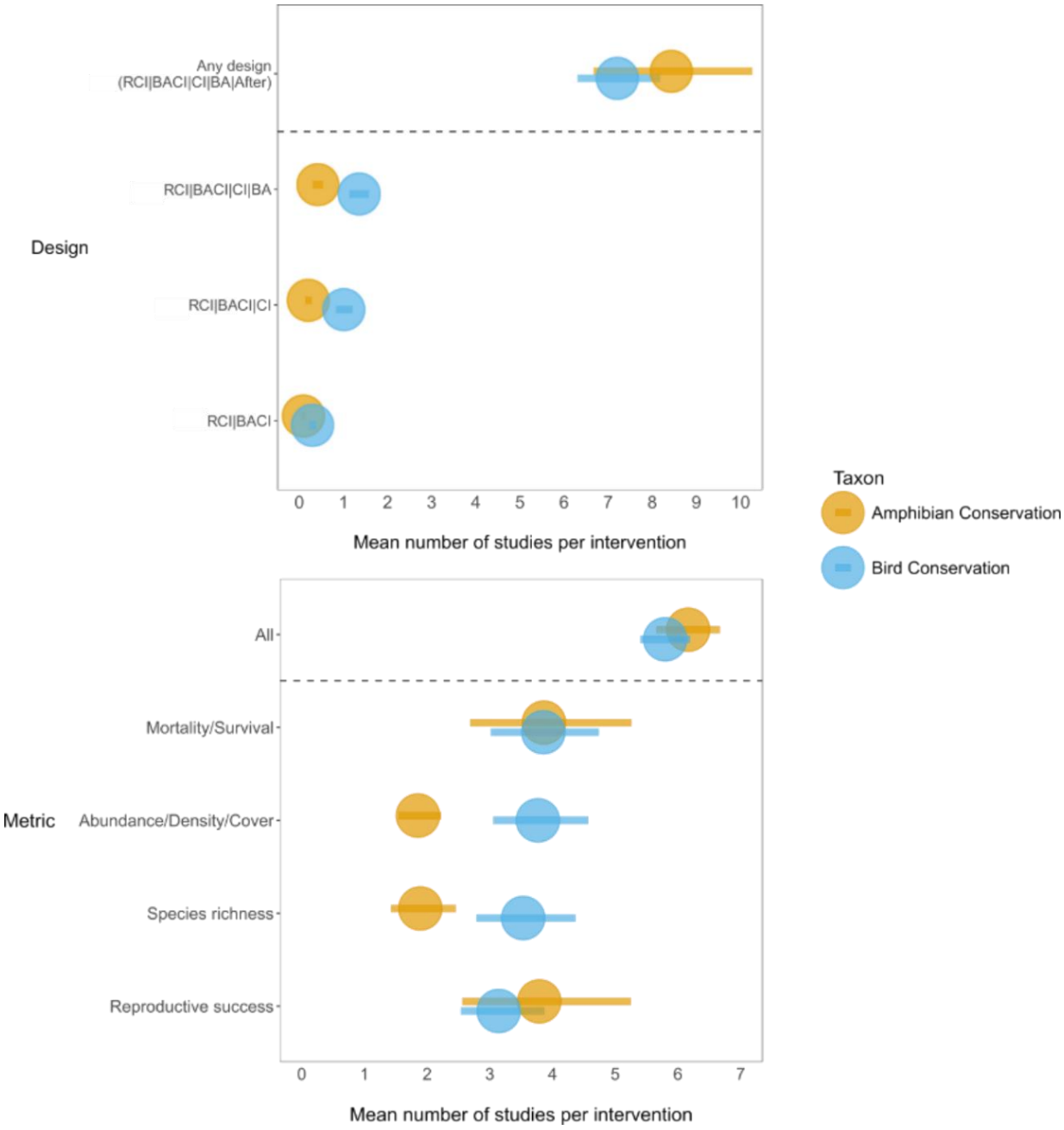


Figure S3 *– Mean number of studies per intervention with different designs and effectiveness metrics. For the top panel, 'Any Design' refers to the mean number of studies per intervention using any of the study designs. Other categories show how this mean value changes when excluding studies using certain designs. | symbolises 'or' – for example, BACI|RCI means studies with BACI or RCI designs. For the bottom panel, 'All' refers to the mean number of studies per intervention using any of the four groups of metrics. Error bars show bootstrapped 95% Confidence Intervals.*

Table S1

Table S1 *– Groupings of similar interventions for amphibians and birds.*

Amphibians	
GroupedInterventionID	**Original_intervention_name**
1	Head-start amphibians for release
2	Translocate amphibians
2	Translocate frogs
2	Translocate great crested newts
2	Translocate natterjack toads
2	Translocate salamanders (including newts)
2	Translocate toads
2	Translocate wood frogs
3	Release captive-bred amphibians
3	Release captive-bred frogs
3	Release captive-bred green and golden bell frogs
3	Release captive-bred Mallorcan midwife toads
3	Release captive-bred salamanders (including newts)
3	Release captive-bred toads
4	Freeze sperm or eggs for future use
5	Use artificial fertilization in captive breeding
6	Use hormone treatment to induce sperm and egg release during captive breeding
7	Captive breeding frogs
7	Captive breeding harlequin toads (Atelopus species)
7	Captive breeding Mallorcan midwife toads
7	Captive breeding salamanders (including newts)
7	Captive breeding toads
8	Legal protection of species
9	Use legislative regulation to protect wild populations
10	Protect habitats for amphibians
11	Reduce impact of amphibian trade
12	Pay farmers to cover the costs of conservation measures
13	Engage landowners and other volunteers to manage land for amphibians
14	Engage volunteers to collect amphibian data (citizen science)
15	Provide education programmes about amphibians
15	Raise awareness amongst the general public through campaigns and public information
16	Exclude fish with barriers
17	Remove or control fish by drying out ponds
18	Remove or control fish by catching
19	Remove or control fish using rotenone
20	Remove or control invasive bullfrogs
21	Remove or control invasive Cuban tree frogs
22	Remove or control viperine snakes
23	Exclude domestic animals or wild hogs by fencing
24	Remove or control mammals
25	Reduce competition from native amphibians

GroupedInterventionID	Original_intervention_name
26	Immunize amphibians against chytridiomycosis infection
27	Use antifungal skin bacteria or peptides to reduce chytridiomycosis infection
28	Use antibacterial treatment to reduce chytridiomycosis infection
29	Use antifungal treatment to reduce chytridiomycosis infection
30	Treat amphibians with chytridiomycosis in the wild or pre-release
31	Use temperature treatment to reduce chytridiomycosis infection
32	Remove the chytrid fungus from ponds
33	Sterilize equipment when moving between amphibian sites
34	Use gloves to handle amphibians
35	Add salt to ponds to reduce chytridiomycosis
36	Install culverts or tunnels as road crossings
37	Install barrier fencing along roads
38	Use humans to assist migrating amphibians across roads
39	Close roads during seasonal amphibian migration
40	Use signage to warn motorists
41	Retain riparian buffer strips during timber harvest
42	Plant riparian buffer strips
43	Restore habitat connectivity
44	Retain connectivity between habitat patches
45	Retain buffer zones around core habitat
46	Mechanically remove mid-storey or ground vegetation
47	Use herbicides to control mid-storey or ground vegetation
48	Clear vegetation
49	Control invasive plants
50	Remove tree canopy to reduce pond shading
51	Use prescribed fire or modifications to burning regime in forests
52	Use prescribed fire or modifications to burning regime in grassland
53	Use leave-tree harvesting instead of clearcutting
54	Use patch retention instead of clearcutting
55	Use shelterwood harvesting instead of clearcutting
56	Thin trees within forests
57	Harvest groups of trees instead of clearcutting
58	Replant vegetation
59	Restore wetland
59	Create wetland
60	Manage ditches
61	Regulate water levels
62	Add lime to water bodies to reduce acidification
63	Create ponds for amphibians
63	Create ponds for frogs
63	Create ponds for great crested newts
63	Create ponds for green toads
63	Create ponds for natterjack toads
63	Create ponds for salamanders (including newts)
63	Create ponds for toads
64	Restore ponds

GroupedInterventionID	Original_intervention_name
65	Deepen, de-silt or re-profile ponds
66	Artificially mist habitat to keep it damp
67	Create refuges
67	Create artificial hibernacula or aestivation sites
68	Leave coarse woody debris in forests
69	Leave standing deadwood/snags in forests
70	Manage grazing regime
71	Change mowing regime
72	Reduce pesticide, herbicide or fertilizer use
73	Create walls or barriers to exclude pollutants
74	Modify gully pots and kerbs
Birds	
GroupedInterventionID	**Original_intervention_name**
1	Alter artificial nest sites to discourage brood parasitism
2	Reduce nest ectoparasites by providing beneficial nesting material
3	Remove/treat endoparasites and diseases
4	Use false brood parasite eggs to discourage brood parasitism
5	Remove brood parasite eggs from target species nests
6	Remove ectoparasites from feathers to increase survival or reproductive success
7	Remove ectoparasites from nests to increase survival or reproductive success
8	Clean nest boxes to increase occupancy or reproductive success
9	Artificially incubate and hand-rear bustards in captivity
9	Artificially incubate and hand-rear cranes in captivity
9	Artificially incubate and hand-rear gamebirds in captivity
9	Artificially incubate and hand-rear parrots in captivity
9	Artificially incubate and hand-rear penguins in captivity
9	Artificially incubate and hand-rear rails in captivity
9	Artificially incubate and hand-rear raptors in captivity
9	Artificially incubate and hand-rear seabirds in captivity
9	Artificially incubate and hand-rear songbirds in captivity
9	Artificially incubate and hand-rear storks and ibises in captivity
9	Artificially incubate and hand-rear vultures in captivity
9	Artificially incubate and hand-rear waders in captivity
9	Artificially incubate and hand-rear wildfowl in captivity
9	Artificially incubate eggs or warm nests
10	Use artificial visual and auditory stimuli to induce breeding in wild populations
11	Foster eggs or chicks of bustards with wild conspecifics
11	Foster eggs or chicks of cranes with wild conspecifics
11	Foster eggs or chicks of cranes with wild non-conspecifics (cross-fostering)
11	Foster eggs or chicks of gannets and boobies with wild conspecifics
11	Foster eggs or chicks of ibises with wild non-conspecifics (cross-fostering)
11	Foster eggs or chicks of owls with wild conspecifics
11	Foster eggs or chicks of parrots with wild conspecifics
11	Foster eggs or chicks of petrels and shearwaters with wild non-conspecifics (cross-fostering)
11	Foster eggs or chicks of raptors with wild conspecifics

GroupedInterventionID	Original_intervention_name
11	Foster eggs or chicks of songbirds with wild non-conspecifics (cross-fostering)
11	Foster eggs or chicks of vultures with wild conspecifics
11	Foster eggs or chicks of waders with wild conspecifics
11	Foster eggs or chicks of waders with wild non-conspecifics (cross-fostering)
11	Foster eggs or chicks of woodpeckers with wild conspecifics
13	Provide artificial nesting sites for burrow-nesting seabirds
13	Provide artificial nesting sites for divers/loons
13	Provide artificial nesting sites for falcons
13	Provide artificial nesting sites for gamebirds
13	Provide artificial nesting sites for grebes
13	Provide artificial nesting sites for ground and tree-nesting seabirds
13	Provide artificial nesting sites for ibises and flamingos
13	Provide artificial nesting sites for oilbirds
13	Provide artificial nesting sites for owls
13	Provide artificial nesting sites for parrots
13	Provide artificial nesting sites for pigeons
13	Provide artificial nesting sites for rails
13	Provide artificial nesting sites for raptors
13	Provide artificial nesting sites for rollers
13	Provide artificial nesting sites for songbirds
13	Provide artificial nesting sites for swifts
13	Provide artificial nesting sites for trogons
13	Provide artificial nesting sites for waders
13	Provide artificial nesting sites for wildfowl
13	Provide artificial nesting sites for wildfowl using artificial/floating islands
13	Provide artificial nesting sites for woodpeckers
14	Provide nesting habitat for birds that is safe from extreme weather
15	Provide nesting material for wild birds
16	Water nesting mounds to increase incubation success in malleefowl
17	Provide bird feeding materials to families with young children
18	Provide calcium supplements to increase survival or reproductive success
19	Provide food for vultures to reduce mortality from diclofenac
20	Can supplementary feeding increase predation or parasitism?
20	Provide supplementary food
20	Provide supplementary food after release
20	Provide supplementary food for auks to increase reproductive success
20	Provide supplementary food for cranes to increase adult survival
20	Provide supplementary food for gamebirds to increase adult survival
20	Provide supplementary food for gamebirds to increase reproductive success
20	Provide supplementary food for gannets and boobies to increase reproductive success
20	Provide supplementary food for gulls, terns and skuas to increase adult survival
20	Provide supplementary food for gulls, terns and skuas to increase reproductive success
20	Provide supplementary food for hummingbirds to increase adult survival
20	Provide supplementary food for ibises to increase reproductive success
20	Provide supplementary food for kingfishers to increase reproductive success

GroupedInterventionID	Original_intervention_name
20	Provide supplementary food for nectar-feeding songbirds to increase adult survival
20	Provide supplementary food for owls to increase reproductive success
20	Provide supplementary food for parrots to increase reproductive success
20	Provide supplementary food for petrels to increase reproductive success
20	Provide supplementary food for pigeons to increase adult survival
20	Provide supplementary food for pigeons to increase reproductive success
20	Provide supplementary food for rails and coots to increase reproductive success
20	Provide supplementary food for raptors to increase adult survival
20	Provide supplementary food for raptors to increase reproductive success
20	Provide supplementary food for songbirds to increase adult survival
20	Provide supplementary food for songbirds to increase reproductive success
20	Provide supplementary food for vultures to increase adult survival
20	Provide supplementary food for vultures to increase reproductive success
20	Provide supplementary food for waders to increase adult survival
20	Provide supplementary food for waders to increase reproductive success
20	Provide supplementary food for wildfowl to increase adult survival
20	Provide supplementary food for wildfowl to increase reproductive success
20	Provide supplementary food for woodpeckers to increase adult survival
20	Provide supplementary food through the establishment of food populations
20	Provide supplementary food to allow the rescue of a second chick
21	Provide supplementary water to increase survival or reproductive success
22	Use perches to increase foraging success
23	Use decoys to attract birds to safe areas
24	Use flashing lights to reduce mortality from artificial lights
25	Use high-visibility mesh on gillnets to reduce seabird bycatch
26	Use lights low in spectral red to reduce mortality from artificial lights
27	Use perch-deterrents to stop raptors perching on pylons
28	Use shark liver oil to reduce seabird bycatch
29	Use repellents to deter birds from landing on pools polluted by mining
30	Use visual and acoustic scarers to deter birds from landing on pools polluted by mining or sewage
31	Use vocalisations to attract birds to safe areas
32	Weight baits or lines to reduce longline bycatch of seabirds
33	Add perches to electricity pylons to reduce electrocution
34	Bury or isolate power lines to reduce incidental bird mortality
35	Insulate power pylons to prevent electrocution
36	Use streamer lines to reduce seabird bycatch on longlines
37	Angle windows to reduce collisions by birds
38	Place feeders close to windows to reduce collisions
39	Deter birds from landing on shellfish culture gear by suspending oyster bags under water
40	Deter birds from landing on shellfish culture gear using spikes on oyster cages
41	Dye baits to reduce seabird bycatch
42	Mark fences to reduce bird collision mortality
43	Mark or tint windows to reduce collision mortality
44	Mark power lines to reduce incidental bird mortality
45	Mark trawler warp cables to reduce seabird collisions

GroupedInterventionID	Original_intervention_name
46	Paint wind turbines to increase their visibility
47	Use raptor models to deter birds and so reduce incidental mortality
48	Reduce electrocutions by using plastic, not aluminium, leg rings to mark birds
49	Set lines underwater to reduce seabird bycatch
50	Remove earth wires to reduce incidental bird mortality
51	Set longlines at night to reduce seabird bycatch
52	Shield lights to reduce mortality from artificial lights
53	Thaw bait before setting lines to reduce seabird bycatch
54	Thicken earth wire to reduce incidental bird mortality
55	Turn deck lights off during night-time setting of longlines to reduce bycatch
56	Turn off lights to reduce mortality from artificial lights
57	Use a line shooter to reduce seabird bycatch
58	Use a sonic scarer when setting longlines to reduce seabird bycatch
59	Use acoustic alerts on gillnets to reduce seabird bycatch
60	Use bait throwers to reduce seabird bycatch
61	Use bird exclusion devices (BEDs) such as Brickle curtains to reduce seabird mortality when hauling longlines
62	Use coloured baits to reduce accidental mortality during predator control
63	Reduce seabird bycatch by releasing offal overboard when setting longlines
64	Reduce conflict by deterring birds from taking crops using bird scarers
65	Scare or otherwise deter birds from airports
66	Scare birds from fish farms
67	Disturb birds at roosts
68	Disturb birds using foot patrols
69	Spray water to deter birds from ponds
70	Alter habitat to encourage birds to leave an area
71	Use electric fencing to exclude fish-eating birds
72	Use mussel socks to prevent birds from attacking shellfish
73	Use netting to exclude fish-eating birds
74	Use in-water devices to reduce fish loss from ponds
75	Increase water turbidity to reduce fish predation by birds
76	Reduce conflict by deterring birds from taking crops using repellents
77	Remove/control adult brood parasites
78	Reduce competition between species by providing nest boxes
79	Reduce inter-specific competition for food by removing or controlling competitor species
80	Reduce inter-specific competition for nest sites by modifying habitats to exclude competitor species
81	Reduce inter-specific competition for nest sites of ground nesting seabirds by removing competitor species
82	Reduce inter-specific competition for nest sites of songbirds by removing competitor species
83	Reduce inter-specific competition for nest sites of woodpeckers by removing competitor species
84	Exclude or control reservoir species to reduce parasite burdens
85	Control avian predators on islands
86	Control invasive ants on islands
87	Control mammalian predators on islands
87	Control mammalian predators on islands for gamebirds

GroupedInterventionID	Original_intervention_name
87	Control mammalian predators on islands for parrots
87	Control mammalian predators on islands for pigeons
87	Control mammalian predators on islands for rails
87	Control mammalian predators on islands for raptors
87	Control mammalian predators on islands for seabirds
87	Control mammalian predators on islands for songbirds
87	Control mammalian predators on islands for waders
87	Control mammalian predators on islands for wildfowl
88	Control or remove habitat-altering mammals
89	Control predators not on islands
89	Control predators not on islands for cranes
89	Control predators not on islands for gamebirds
89	Control predators not on islands for parrots
89	Control predators not on islands for rails
89	Control predators not on islands for seabirds
89	Control predators not on islands for songbirds
89	Control predators not on islands for waders
89	Control predators not on islands for wildfowl
89	Remove or control predators to enhance bird populations and communities
90	Use snakeskin to deter mammalian nest predators
91	Use naphthalene to deter mammalian predators
92	Use repellents on baits for predator control
93	Distribute poison bait for predator control using dispensers
93	Do birds take bait designed for pest control?
94	Reduce adverse habitat alterations by excluding problematic aquatic species
95	Reduce adverse habitat alterations by excluding problematic terrestrial species
96	Can nest protection increase nest abandonment?
96	Can nest protection increase predation of adults and chicks?
96	Physically protect nests from predators using non-electric fencing
96	Physically protect nests with individual exclosures/barriers or provide shelters for chicks of ground nesting seabirds
96	Physically protect nests with individual exclosures/barriers or provide shelters for chicks of songbirds
96	Physically protect nests with individual exclosures/barriers or provide shelters for chicks of storks and ibises
96	Physically protect nests with individual exclosures/barriers or provide shelters for chicks of waders
96	Protect nests from livestock to reduce trampling
96	Reduce nest predation by excluding predators from nests or nesting areas
97	Guard nests to increase nest success
98	Increase on-the-ground protection to reduce unsustainable levels of exploitation
99	Introduce voluntary maximum shoot distances
100	Mark nests during harvest
101	Mark eggs to reduce their appeal to egg collectors
102	Protect bird nests using electric fencing
103	Protect nest sites from competitors
104	Protect nests from ants
105	Provide paths to limit the extent of disturbance

GroupedInterventionID	Original_intervention_name
106	Use supplementary feeding to reduce predation
107	Use wildlife refuges to reduce hunting disturbance
108	Reduce predation by translocating nest boxes
109	Reduce predation by translocating predators
110	Use artificial nests that discourage predation
111	Use differently-coloured artificial nests
112	Use aversive conditioning to reduce nest predation by avian predators
113	Use aversive conditioning to reduce nest predation by mammalian predators
114	Use collar-mounted devices to reduce predation
115	Use copper strips to exclude snails from nests
116	Remove eggs from wild nests to increase reproductive output
117	Repair/support nests to support breeding
118	Replace nesting substrate following severe weather
119	Use mowing techniques to reduce chick mortality
120	Use multiple barriers to protect nests
121	Use nest covers to reduce the impact of research on predation of ground-nesting seabirds
122	Habituate birds to human visitors
123	Use voluntary agreements with local people to reduce disturbance
124	Use signs and access restrictions to reduce disturbance at nest sites
125	Raise awareness amongst the general public through campaigns and public information
125	Use education programmes and local engagement to help reduce persecution or exploitation of species
126	Start educational programmes for personal watercraft owners
127	Employ local people as biomonitors
128	Legally protect habitats
129	Offer per clutch payment for farmland birds
130	Pay farmers to cover the costs of bird conservation measures
131	Use legislative regulation to protect wild populations
132	Wash contaminated semen and use it for artificial insemination
133	Freeze semen for use in artificial insemination
134	Use artificial insemination in captive breeding
135	Can captive breeding have deleterious effects on individual fitness?
135	Use captive breeding to increase or maintain populations of bustards
135	Use captive breeding to increase or maintain populations of cranes
135	Use captive breeding to increase or maintain populations of pigeons
135	Use captive breeding to increase or maintain populations of rails
135	Use captive breeding to increase or maintain populations of raptors
135	Use captive breeding to increase or maintain populations of seabirds
135	Use captive breeding to increase or maintain populations of songbirds
135	Use captive breeding to increase or maintain populations of storks and ibises
135	Use captive breeding to increase or maintain populations of tinamous
136	Use appropriate populations to source released populations
137	Use microlites to help birds migrate
138	Use flying training before release
139	Use anti-predator training to improve survival after release

GroupedInterventionID	Original_intervention_name
140	Ensure translocated birds are familiar with each other before release
141	Release birds as adults or sub-adults, not juveniles
142	Release birds in coveys
143	Release birds in groups
144	Release captive-bred individuals into the wild to restore or augment wild populations of bustards
144	Release captive-bred individuals into the wild to restore or augment wild populations of cranes
144	Release captive-bred individuals into the wild to restore or augment wild populations of gamebirds
144	Release captive-bred individuals into the wild to restore or augment wild populations of owls
144	Release captive-bred individuals into the wild to restore or augment wild populations of parrots
144	Release captive-bred individuals into the wild to restore or augment wild populations of pigeons
144	Release captive-bred individuals into the wild to restore or augment wild populations of rails
144	Release captive-bred individuals into the wild to restore or augment wild populations of raptors
144	Release captive-bred individuals into the wild to restore or augment wild populations of songbirds
144	Release captive-bred individuals into the wild to restore or augment wild populations of storks and ibises
144	Release captive-bred individuals into the wild to restore or augment wild populations of vultures
144	Release captive-bred individuals into the wild to restore or augment wild populations of waders
144	Release captive-bred individuals into the wild to restore or augment wild populations of wildfowl
145	Relocate birds following oil spills
146	Relocate nestlings to reduce poaching
147	Relocate nests at harvest time to reduce nestling mortality
148	Use holding pens at release sites
149	Translocate auks
149	Translocate birds away from fish farms
149	Translocate gamebirds
149	Translocate herons, storks and ibises
149	Translocate individuals
149	Translocate megapodes
149	Translocate nests to avoid disturbance
149	Translocate owls
149	Translocate parrots
149	Translocate pelicans
149	Translocate petrels and shearwaters
149	Translocate rails
149	Translocate raptors
149	Translocate songbirds
149	Translocate wildfowl
149	Translocate woodpeckers
150	Move fish-eating birds to reduce conflict with fishermen
151	Use puppets to increase the survival or growth of hand-reared chicks
152	Use techniques to increase the survival of species after capture

GroupedInterventionID	Original_intervention_name
153	Clip birds wings on release
154	Rehabilitation of injured and treated birds
155	Clean birds following oil spills
156	Add woody debris to forests
157	Create beetle banks
158	Clear or open patches in forests
159	Clearcut and re-seed forests
160	Manually control or remove midstorey and ground-level vegetation (including mowing, chaining, cutting etc) in forests
160	Manually control or remove midstorey and ground-level vegetation (including mowing, chaining, cutting etc) in shrubland
161	Apply herbicide to mid- and understorey vegetation
162	Coppice trees
163	Control scrub on farmland
164	Convert to or maintain organic farming systems
165	Delay haying/mowing
166	Leave headlands in fields unsprayed (conservation headlands)
167	Exclude grazers from semi-natural habitats
168	Create scrapes and pools in wetlands and wet grasslands
169	Create skylark plots for bird conservation
170	Create uncultivated margins around intensive arable or pasture fields for birds
171	Employ grazing in artificial grasslands/pastures
171	Employ grazing in natural grasslands
171	Employ grazing in non-grassland habitats
172	Create open patches or strips in permanent grassland
173	Mow or cut natural grasslands
173	Mow or cut semi-natural grasslands/pastures
173	Mowing roadside verges
174	Mow or cut reedbeds
175	Use traditional breeds of livestock
176	Plough habitats
177	Increase the proportion of natural/semi-natural vegetation in the farmed landscape
178	Increase crop diversity to benefit birds
179	Ensure connectivity between habitat patches
180	Leave overwinter stubbles
181	Fertilize artificial grasslands
182	Leave refuges in fields during harvest
182	Leave uncropped, cultivated margins or plots, including lapwing and stone curlew plots
182	Provide or retain set-aside areas in farmland
182	Provide or retain un-harvested buffer strips
183	Leave uncut rye grass in silage fields for birds
184	Maintain species-rich, semi-natural grassland
185	Maintain traditional orchards
186	Maintain traditional water meadows
187	Maintain upland heath/moor
188	Manage ditches to benefit wildlife

GroupedInterventionID	Original_intervention_name
189	Manage hedges to benefit birds
190	Manage water level in wetlands
191	Manage woodland edges for birds
192	Plant cereals for whole crop silage
193	Plant cereals in wide-spaced rows
194	Plant grass buffer strips/margins around arable or pasture fields for birds
195	Plant more than one crop per field (intercropping)
196	Plant nectar flower mixture/wildflower strips for birds
197	Plant new hedges
198	Plant trees to act as windbreaks
199	Plant wild bird seed or cover mixture
200	Promote sustainable alternative livelihoods
201	Provide refuges for fish within ponds
202	Provide sacrificial grasslands to reduce the impact of wild geese on crops
203	Provide short grass for waders
204	Raise mowing height on grasslands to benefit birds
205	Raise water levels in ditches or grassland
206	Reduce chemical inputs in permanent grassland management
207	Reduce grazing intensity
208	Reduce management intensity on permanent grasslands for birds
209	Treat wetlands with herbicide
210	Reduce pesticide or herbicide use generally
211	Reduce tillage
212	Use buffer zones to reduce the impact of invasive plant control
213	Remove coarse woody debris from forests
214	Remove midstorey from savannas
215	Remove problematic vegetation
216	Remove vegetation to create nesting areas
217	Replace non-native species of tree/shrub
218	Re-seed grasslands
219	Restore or create coastal and intertidal wetlands
220	Restore or create forests
221	Restore or create grasslands
222	Restore or create inland wetlands
223	Restore or create kelp forests
224	Restore or create lagoons
225	Restore or create shrubland
226	Restore or create traditional water meadows
227	Restrict certain pesticides or other agricultural chemicals for birds
228	Revert arable land to permanent grassland
229	Sow crops in spring rather than autumn
230	Take field corners out of management
231	Undersow spring cereals, with clover for example
232	Use lime to reduce acidification in lakes
233	Use environmentally sensitive flood management
234	Use fire suppression/control

GroupedInterventionID	Original_intervention_name
235	Use greentree reservoir management
236	Use mosaic management
237	Use prescribed burning on Australian sclerophyll forest
237	Use prescribed burning on coastal habitats
237	Use prescribed burning on deciduous forests
237	Use prescribed burning on grasslands
237	Use prescribed burning on pine forests
237	Use prescribed burning on savannas
237	Use prescribed burning on shrublands
238	Use variable retention management during forestry operations
239	Use selective harvesting/logging instead of clearcutting
240	Use patch retention harvesting instead of clearcutting
241	Thin trees within forests
242	Use shelterwood cutting instead of clearcutting
243	Use ring-barking (girdling), cutting or silvicides to produce snags

Table S2

Table S2 *– Records of accuracy checks of the study design detected for a random subset of studies (approximately 5% of the total number of studies for each synopsis). ID = unique study identifier; Correct = whether the correct study design was identified; Detected_design = if there was an error, which design was detected?; True_design = if there was an error, what was the true design based on the original text of study?; Synopsis = which synopsis was the study in?; Intervention_name_original = Original name of intervention (matched to website – see Table S1 to match to grouped interventions we used).*

Amphibians					
ID	**Correct**	**Detected_design**	**True_design**	**Synopsis**	**Intervention_name_original**
4828	Y	NA	NA	Amphibian Conservation	Restore ponds
4828	Y	NA	NA	Amphibian Conservation	Manage grazing regime
4845	Y	NA	NA	Amphibian Conservation	Captive breeding toads
4845	Y	NA	NA	Amphibian Conservation	Use hormone treatment to induce sperm and egg release during captive breeding
4869	Y	NA	NA	Amphibian Conservation	Create ponds for natterjack toads
4869	Y	NA	NA	Amphibian Conservation	Translocate natterjack toads
4869	Y	NA	NA	Amphibian Conservation	Add lime to water bodies to reduce acidification
4878	Y	NA	NA	Amphibian Conservation	Use antifungal treatment to reduce chytridiomycosis infection
4895	Y	NA	NA	Amphibian Conservation	Replant vegetation
4963	Y	NA	NA	Amphibian Conservation	Clear vegetation
4963	Y	NA	NA	Amphibian Conservation	Manage grazing regime
4966	Y	NA	NA	Amphibian Conservation	Translocate great crested newts
4966	Y	NA	NA	Amphibian Conservation	Clear vegetation
4996	Y	NA	NA	Amphibian Conservation	Create ponds for great crested newts
5012	Y	NA	NA	Amphibian Conservation	Translocate toads
5012	Y	NA	NA	Amphibian Conservation	Head-start amphibians for release
5018	N	BA	After	Amphibian Conservation	Create wetland
5100	Y	NA	NA	Amphibian Conservation	Remove or control invasive bullfrogs
5107	Y	NA	NA	Amphibian Conservation	Engage landowners and other volunteers to manage land for amphibians
5119	Y	NA	NA	Amphibian Conservation	Captive breeding Mallorcan midwife toads
5124	Y	NA	NA	Amphibian Conservation	Use antifungal treatment to reduce chytridiomycosis infection
Birds					
ID	**Correct**	**Detected_design**	**True_design**	**Synopsis**	**Intervention_name_original**
222	Y	NA	NA	Bird Conservation	Control mammalian predators on islands for seabirds
1137	Y	NA	NA	Bird Conservation	Provide artificial nesting sites for songbirds

ID	Correct	Detected_design	True_design	Synopsis	Intervention_name_original
1195	N	After	CI	Bird Conservation	Provide artificial nesting sites for songbirds
1281	Y	NA	NA	Bird Conservation	Remove/treat endoparasites and diseases
1311	Y	NA	NA	Bird Conservation	Translocate songbirds
1321	Y	NA	NA	Bird Conservation	Provide supplementary food for songbirds to increase adult survival
1322	Y	NA	NA	Bird Conservation	Provide supplementary food for waders to increase reproductive success
1325	Y	NA	NA	Bird Conservation	Provide supplementary food for gulls, terns and skuas to increase reproductive success
1331	Y	NA	NA	Bird Conservation	Provide artificial nesting sites for songbirds
1345	Y	NA	NA	Bird Conservation	Provide supplementary food for gannets and boobies to increase reproductive success
1364	Y	NA	NA	Bird Conservation	Use education programmes and local engagement to help reduce persecution or exploitation of species
1364	Y	NA	NA	Bird Conservation	Release birds as adults or sub-adults, not juveniles
1376	Y	NA	NA	Bird Conservation	Mow or cut natural grasslands
1394	Y	NA	NA	Bird Conservation	Foster eggs or chicks of woodpeckers with wild conspecifics
1432	Y	NA	NA	Bird Conservation	Provide artificial nesting sites for wildfowl
1469	Y	NA	NA	Bird Conservation	Provide artificial nesting sites for ground and tree-nesting seabirds
1523	Y	NA	NA	Bird Conservation	Restore or create forests
1574	Y	NA	NA	Bird Conservation	Provide supplementary food for vultures to increase adult survival
1574	Y	NA	NA	Bird Conservation	Restrict certain pesticides or other agricultural chemicals for birds
1574	Y	NA	NA	Bird Conservation	Use legislative regulation to protect wild populations
1583	Y	NA	NA	Bird Conservation	Release birds as adults or sub-adults, not juveniles
1646	Y	NA	NA	Bird Conservation	Remove/control adult brood parasites
1684	Y	NA	NA	Bird Conservation	Reduce seabird bycatch by releasing offal overboard when setting longlines
1684	Y	NA	NA	Bird Conservation	Set longlines at night to reduce seabird bycatch
1684	Y	NA	NA	Bird Conservation	Turn deck lights off during night-time setting of longlines to reduce bycatch
1702	Y	NA	NA	Bird Conservation	Remove/control adult brood parasites
1788	Y	NA	NA	Bird Conservation	Use prescribed burning on coastal habitats
1934	Y	NA	NA	Bird Conservation	Exclude grazers from semi-natural habitats
2075	Y	NA	NA	Bird Conservation	Use netting to exclude fish-eating birds
2083	Y	NA	NA	Bird Conservation	Disturb birds using foot patrols
2091	Y	NA	NA	Bird Conservation	Use visual and acoustic scarers to deter birds from landing on pools polluted by mining or sewage
2113	Y	NA	NA	Bird Conservation	Use prescribed burning on savannas

ID	Correct	Detected_design	True_design	Synopsis	Intervention_name_original
2116	Y	NA	NA	Bird Conservation	Use prescribed burning on pine forests
2119	N	CI	BACI	Bird Conservation	Thin trees within forests
2143	Y	NA	NA	Bird Conservation	Employ grazing in artificial grasslands/pastures
2229	Y	NA	NA	Bird Conservation	Restore or create traditional water meadows
2252	Y	NA	NA	Bird Conservation	Physically protect nests with individual exclosures/barriers or provide shelters for chicks of waders
2368	Y	NA	NA	Bird Conservation	Raise water levels in ditches or grassland
2370	Y	NA	NA	Bird Conservation	Manage water level in wetlands
2428	Y	NA	NA	Bird Conservation	Manually control or remove midstorey and ground-level vegetation (including mowing, chaining, cutting etc) in shrubland
2428	Y	NA	NA	Bird Conservation	Apply herbicide to mid- and understorey vegetation
3134	Y	NA	NA	Bird Conservation	Use streamer lines to reduce seabird bycatch on longlines
3146	Y	NA	NA	Bird Conservation	Translocate gamebirds
3270	Y	NA	NA	Bird Conservation	Reduce adverse habitat alterations by excluding problematic terrestrial species
3270	Y	NA	NA	Bird Conservation	Restore or create forests
3270	Y	NA	NA	Bird Conservation	Exclude grazers from semi-natural habitats
3338	Y	NA	NA	Bird Conservation	Provide supplementary food for songbirds to increase reproductive success
3357	Y	NA	NA	Bird Conservation	Manually control or remove midstorey and ground-level vegetation (including mowing, chaining, cutting etc) in forests
3420	Y	NA	NA	Bird Conservation	Control predators not on islands for wildfowl
3560	Y	NA	NA	Bird Conservation	Plant new hedges
3560	Y	NA	NA	Bird Conservation	Reduce grazing intensity
3560	Y	NA	NA	Bird Conservation	Restore or create grasslands
3603	Y	NA	NA	Bird Conservation	Use streamer lines to reduce seabird bycatch on longlines
3603	Y	NA	NA	Bird Conservation	Weight baits or lines to reduce longline bycatch of seabirds
3643	Y	NA	NA	Bird Conservation	Control mammalian predators on islands for gamebirds
3672	N	CI	RCI	Bird Conservation	Remove ectoparasites from nests to increase survival or reproductive success
3677	Y	NA	NA	Bird Conservation	Use false brood parasite eggs to discourage brood parasitism
3753	Y	NA	NA	Bird Conservation	Provide supplementary food for songbirds to increase reproductive success
3754	N	CI	RCI	Bird Conservation	Provide supplementary food for songbirds to increase reproductive success
3768	Y	NA	NA	Bird Conservation	Provide supplementary food for gamebirds to increase adult survival
3778	Y	NA	NA	Bird Conservation	Provide supplementary food through the establishment of food populations
3835	Y	NA	NA	Bird Conservation	Artificially incubate and hand-rear songbirds in captivity

Table S3

Table S3 *– Records of accuracy checks of the metrics detected for a random subset of studies (approximately 5% of the total number of studies for each synopsis). ID = unique study identifier; Metric_group_correct = whether the correct metric group was identified; Metric_group_error = if there was an error, which metric group was it for?; False_pos = was it a false positive?; False_neg = was it a false negative?*

Amphibians				
ID	**Metric_group_correct**	**Metric_group_error**	**False_pos**	**False_neg**
4799	1	NA	NA	NA
4814	1	NA	NA	NA
4872	1	NA	NA	NA
4894	1	NA	NA	NA
4895	1	NA	NA	NA
4900	1	NA	NA	NA
4916	1	NA	NA	NA
4925	1	NA	NA	NA
4942	1	NA	NA	NA
4973	1	NA	NA	NA
4989	1	NA	NA	NA
4995	0	Abundance, density, and cover	0	1
5003	1	NA	NA	NA
5013	1	NA	NA	NA
5053	1	NA	NA	NA
5078	1	NA	NA	NA
5114	0	Abundance, density, and cover	0	1
5123	1	NA	NA	NA
5142	1	NA	NA	NA
5158	1	NA	NA	NA
5268	1	NA	NA	NA
Birds				
ID	**Metric_group_correct**	**Metric_group_error**	**False_pos**	**False_neg**
43	1	NA	NA	NA
110	1	NA	NA	NA
492	1	NA	NA	NA
691	1	NA	NA	NA
717	1	NA	NA	NA
781	1	NA	NA	NA
828	0	Survival	1	0
1167	1	NA	NA	NA
1229	1	NA	NA	NA
1305	1	NA	NA	NA
1330	1	NA	NA	NA
1336	1	NA	NA	NA
1355	1	NA	NA	NA
1363	1	NA	NA	NA
1389	1	NA	NA	NA

ID	Metric_group_correct	Metric_group_error	False_pos	False_neg
1423	1	NA	NA	NA
1450	1	NA	NA	NA
1454	1	NA	NA	NA
1466	1	NA	NA	NA
1494	1	NA	NA	NA
1503	1	NA	NA	NA
1583	1	NA	NA	NA
1586	1	NA	NA	NA
1686	1	NA	NA	NA
1898	1	NA	NA	NA
2072	1	NA	NA	NA
2076	1	NA	NA	NA
2091	1	NA	NA	NA
2096	1	NA	NA	NA
2135	1	NA	NA	NA
2236	1	NA	NA	NA
2245	1	NA	NA	NA
2259	1	NA	NA	NA
2377	1	NA	NA	NA
2406	1	NA	NA	NA
2411	1	NA	NA	NA
2413	1	NA	NA	NA
2457	1	NA	NA	NA
2859	1	NA	NA	NA
3159	1	NA	NA	NA
3161	1	NA	NA	NA
3198	1	NA	NA	NA
3209	1	NA	NA	NA
3273	1	NA	NA	NA
3302	1	NA	NA	NA
3330	1	NA	NA	NA
3343	1	NA	NA	NA
3440	0	Reproductive success	1	0
3526	0	Diversity and species richness	0	1
3640	1	NA	NA	NA
3647	1	NA	NA	NA
3648	1	NA	NA	NA
3664	1	NA	NA	NA
3669	1	NA	NA	NA
3676	1	NA	NA	NA
3707	1	NA	NA	NA
3742	1	NA	NA	NA
3815	1	NA	NA	NA
3822	1	NA	NA	NA
3829	1	NA	NA	NA
3830	1	NA	NA	NA
3835	1	NA	NA	NA

Table S4

Table S4 *– Akaike's Information Criterion (AIC) and quasi-R^2 values used in model selection process for binomial Generalised Linear Models with a logistic link. The spatial relationship between the number of studies and the number of species, threatened species and data-deficient species within 2x2-degree grid cells was tested by three separate models for amphibians and birds. All untransformed models were selected and are presented in Table S6.*

Taxon	**Model**	**Transformation**	**AIC**	**quasi-R^2**
Amphibians	Species	None	79.7	0.01
		Square root	79.3	0.01
	Threatened species	None	68.8	0.22
		Square root	68.7	0.24
	Data-deficient species	None	68.9	0.41
		Square root	68.9	0.41
Birds	Species	None	87.0	0.04
		Square root	86.6	0.08
	Threatened species	None	84.7	0.07
		Square root	85.3	0.07
	Data-deficient species	None	66.2	0.43
		Square root	66.2	0.44

Table S5

Table S5 *– Number of countries where at least one study using a given study design was present and continents where no studies using a given design were present (see Table 1 for details of designs). The 'Continents unrepresented' column for amphibians excludes Antarctica as no amphibian species occur there.*

Design	Design acronym	Amphibians		Birds	
		No. of countries represented	**Continents unrepresented**	**No. of countries represented**	**Continents unrepresented**
After		31	None	46	Antarctica
Before-After	BA	23	South America	37	Antarctica
Control-Impact	CI	18	None	38	None
Before-After Control-Impact	BACI	10	South America, Africa, Australasia	20	Antarctica
Randomised Control-Impact	RCI	5	South America, Africa, Asia	15	None

Table S6

Table S6 *– Results of binomial Generalised Linear Models with a logistic link testing the spatial relationship between the number of studies and the number of species, threatened species and data-deficient species within 2x2-degree grid cells in three separate models for amphibians and birds. Coefficients and standard errors are in log odds (3.s.f.). p-values with an asterisk are statistically significant (p<0.05). All models are untransformed (see Table S4 for description of AIC and quasi-R^2 values used for model selection).*

Taxon	**Model**	**Intercept**		**Number of species**		
		Coef.	**SE**	**Coef.**	**SE**	**p-value**
Amphibians	Species	-4.72	0.257	-1.67	1.86	0.37
	Threatened species	-2.76	0.221	-12.6	5.70	0.03*
	Data-deficient species	-2.75	0.214	-25.5	12.1	0.04*
Birds	Species	-6.86	0.280	2.79	0.829	0.00*
	Threatened species	-5.20	0.354	-7.42	2.44	0.00*
	Data-deficient species	-2.84	0.221	-12.7	5.94	0.03*

Appendix S1

Appendix S1 *– Code to detect the metric type used by a study that can be pasted into a .R file to run this code in R.*

```
####load packages
debug(utils:::unpackPkgZip)
require("RCurl")
require('tm')
library(doParallel)
require(XML)

#### function to web scrape the conservation evidence website
masterloop <- function(zeta){
  #### website page ID = unique study ID
  z=zeta
  #### download website page text
  st2835 = getURL(paste("https://www.conservationevidence.com/individual-study/",z,sep=""),ssl.verifypeer=FALSE) ### access conservation evidence website and summary of study with ID zeta (e.g., 1, 2, 3 ...)

  st2835tree = htmlTreeParse(st2835,useInternal = TRUE)

  st2835tree.html=do.call(paste, as.list(capture.output(st2835tree)))

  masterlog = NULL
  if(length(grep("<h2>Summary</h2>",st2835tree.html))==0){ ### Only look at website pages that do not have an old-style heading, these are archived
### and not used anymore and are not assgined to an intervention or synopsis

    st2835tree.parse = unlist(xpathApply(st2835tree,path="//p",fun=xmlValue))

    ####### select study paragraphs and remove extra paragraphs that are on website
    if (length(grep("This option allows you to download this individual study.",st2835tree.parse))==0){
      st2835.txt = list()
      j=1
      for (i in 4:(length(st2835tree.parse)-3)) {
        st2835.txt[[j]] = tolower(as.character(st2835tree.parse[i]))
        j=j+1
      }
    }

    if (length(grep("This option allows you to download this individual study.",st2835tree.parse))>0){
      st2835.txt = list()
      j=1
      for (i in 4:(max(grep("This option allows you to download this individual study.",st2835tree.parse))-1)) {
        st2835.txt[[j]] = tolower(as.character(st2835tree.parse[i]))
        j=j+1
      }
    }

    ###### use regular expressions to identify metrics separately (then group later during analysis – see Materials and methods for details)
    masterlog = matrix(ncol=10,nrow=length(st2835.txt))

    for(i in 1:length(st2835.txt)){

      ###density
```

```
    if(length(grep("/plot",st2835.txt[[i]])) > 0 |length(grep("volume of arthropods",st2835.txt[[i]])) > 0
|length(grep("earthworms/m2",st2835.txt[[i]])) > 0 |length(grep(" per fish",st2835.txt[[i]])) > 0
|length(grep("density",st2835.txt[[i]])) > 0 |length(grep("densities",st2835.txt[[i]])) > 0
|length(grep("captures increased",st2835.txt[[i]])) > 0|length(grep("captures
decreased",st2835.txt[[i]])) > 0){masterlog[i,1]=1}
    if(length(grep("/plot",st2835.txt[[i]])) == 0 &length(grep("volume of arthropods",st2835.txt[[i]])) ==
0 &length(grep("earthworms/m2",st2835.txt[[i]])) == 0 &length(grep(" per fish",st2835.txt[[i]])) == 0
&length(grep("density",st2835.txt[[i]])) == 0 &length(grep("densities",st2835.txt[[i]])) == 0
&length(grep("captures increased",st2835.txt[[i]])) == 0&length(grep("captures
decreased",st2835.txt[[i]])) == 0){masterlog[i,1]=0}
    if(length(grep("effect of.{,20}density",st2835.txt[[i]])) > 0 | length(grep("effect
of.{,20}density",st2835.txt[[i]])) > 0){masterlog[i,1]=0}

    ##abundance
    if(length(grep("were counted",st2835.txt[[i]])) > 0 |length(grep("more nest-
searching",st2835.txt[[i]])) > 0 |length(grep("more foraging",st2835.txt[[i]])) > 0 |length(grep("number
of breeding",st2835.txt[[i]])) > 0 |length(grep("more microorganisms",st2835.txt[[i]])) > 0
|length(grep("similar numbers",st2835.txt[[i]])) > 0 |length(grep("more numerous",st2835.txt[[i]])) > 0
|length(grep("numbers of nest-searching",st2835.txt[[i]])) > 0 |length(grep("numbers of
foraging",st2835.txt[[i]])) > 0 |length(grep("pest numbers",st2835.txt[[i]])) > 0 |length(grep("doubled a
population",st2835.txt[[i]])) > 0 |length(grep("population.{,30}doubl",st2835.txt[[i]])) > 0 |length(grep("
more pollinators ",st2835.txt[[i]])) > 0 |length(grep(" more.{,25}were seen ",st2835.txt[[i]])) > 0
|length(grep("counts of",st2835.txt[[i]])) > 0 |length(grep("total number",st2835.txt[[i]])) > 0
|length(grep("abundance",st2835.txt[[i]])) > 0 |length(grep("more abundant",st2835.txt[[i]])) > 0
|length(grep("less abundant",st2835.txt[[i]])) > 0
|length(grep("population.{,50}increase",st2835.txt[[i]])) > 0
|length(grep("population.{,50}decrease",st2835.txt[[i]])) > 0
|length(grep("in.{,25}number",st2835.txt[[i]])) > 0|length(grep("crease
in.{,35}population",st2835.txt[[i]])) > 0|length(grep("numbers.{,20}..creased",st2835.txt[[i]])) >
0|length(grep("higher.{,20}number",st2835.txt[[i]])) >
0|length(grep("lower.{,20}number",st2835.txt[[i]])) > 0){masterlog[i,2]=1}
    if(length(grep("were counted",st2835.txt[[i]])) == 0 &length(grep("more nest-
searching",st2835.txt[[i]])) == 0 &length(grep("more foraging",st2835.txt[[i]])) == 0
&length(grep("number of breeding",st2835.txt[[i]])) == 0 &length(grep("more
microorganisms",st2835.txt[[i]])) == 0 &length(grep("similar numbers",st2835.txt[[i]])) == 0
&length(grep("more numerous",st2835.txt[[i]])) == 0 &length(grep("numbers of nest-
searching",st2835.txt[[i]])) == 0 &length(grep("numbers of foraging",st2835.txt[[i]])) == 0
&length(grep("pest numbers",st2835.txt[[i]])) == 0 &length(grep("doubled a
population",st2835.txt[[i]])) == 0 &length(grep("population.{,30}doubl",st2835.txt[[i]])) == 0
&length(grep(" more pollinators ",st2835.txt[[i]])) == 0 &length(grep(" more.{,25}were seen
",st2835.txt[[i]])) == 0 &length(grep("counts of",st2835.txt[[i]])) == 0 &length(grep("total
number",st2835.txt[[i]])) == 0&length(grep("abundance",st2835.txt[[i]])) == 0 & length(grep("more
abundant",st2835.txt[[i]])) == 0 &length(grep("less abundant",st2835.txt[[i]])) == 0
&length(grep("population.{,50}increase",st2835.txt[[i]])) == 0
&length(grep("population.{,50}decrease",st2835.txt[[i]])) == 0&length(grep("crease
in.{,35}population",st2835.txt[[i]])) == 0&length(grep("numbers.{,20}..creased",st2835.txt[[i]])) ==
0&length(grep("higher.{,20}number",st2835.txt[[i]])) ==
0&length(grep("lower.{,20}number",st2835.txt[[i]])) == 0){masterlog[i,2]=0}

    ##diversity
    if(length(grep(" more.{,25}species ",st2835.txt[[i]])) > 0 |length(grep("richness",st2835.txt[[i]])) > 0
|length(grep(" diversity ",st2835.txt[[i]])) > 0|length(grep(" numbers of species ",st2835.txt[[i]])) > 0
|length(grep(" number of species ",st2835.txt[[i]])) > 0|length(grep(" numbers of.{,25}species
",st2835.txt[[i]])) > 0 |length(grep(" number of.{,25}species ",st2835.txt[[i]])) > 0
|length(grep("community composition",st2835.txt[[i]])) > 0|length(grep("species
composition",st2835.txt[[i]])) > 0){masterlog[i,3]=1}
    if(length(grep(" more.{,25}species ",st2835.txt[[i]])) == 0 &length(grep("richness",st2835.txt[[i]]))
== 0 &length(grep(" diversity ",st2835.txt[[i]])) == 0&length(grep(" numbers of species
",st2835.txt[[i]])) == 0&length(grep(" number of species ",st2835.txt[[i]])) == 0&length(grep(" numbers
of.{,25}species ",st2835.txt[[i]])) == 0&length(grep(" number of.{,25}species ",st2835.txt[[i]])) ==
```

```
0&length(grep("community composition",st2835.txt[[i]])) == 0&length(grep("species composition",st2835.txt[[i]])) == 0){masterlog[i,3]=0}
    if(length(grep(" small numbers of these species ",st2835.txt[[i]])) > 0){masterlog[i,3]=0}

    ##cover
    if(length(grep("effect on.{,15}cover",st2835.txt[[i]])) > 0 |length(grep("herbaceous cover",st2835.txt[[i]])) > 0 |length(grep("percentage.{,15}cover",st2835.txt[[i]])) > 0 |length(grep("% cover",st2835.txt[[i]])) > 0 |length(grep("average.{,15}cover",st2835.txt[[i]])) > 0|length(grep("species cover",st2835.txt[[i]])) > 0|length(grep("greater.{,15}cover",st2835.txt[[i]])) > 0|length(grep("lower.{,15}cover",st2835.txt[[i]])) > 0|length(grep("higher.{,15}cover ",st2835.txt[[i]])) > 0|length(grep("less.{,15}cover",st2835.txt[[i]])) > 0|length(grep("more.{,15}cover",st2835.txt[[i]])) > 0|length(grep(" cover of",st2835.txt[[i]]))>0|length(grep(" cover decreased",st2835.txt[[i]])) > 0|length(grep(" cover increased",st2835.txt[[i]])) > 0|length(grep("decreased.{,15}cover",st2835.txt[[i]])) > 0|length(grep("..creased.{,15}cover",st2835.txt[[i]])) > 0|length(grep("less.{,25}cover",st2835.txt[[i]])) > 0|length(grep("more.{,25}cover",st2835.txt[[i]])) > 0 |length(grep("total.{,25}cover",st2835.txt[[i]])) > 0 | length(grep("proportion of.{,15}cover",st2835.txt[[i]])) > 0 | length(grep("cover was higher",st2835.txt[[i]])) > 0 | length(grep("cover was lower",st2835.txt[[i]])) > 0 | length(grep("% of the surface was cover",st2835.txt[[i]])) > 0 ){masterlog[i,4]=1}
    if(length(grep("effect on.{,15}cover",st2835.txt[[i]])) == 0 &length(grep("herbaceous cover",st2835.txt[[i]])) == 0 &length(grep("percentage.{,15}cover",st2835.txt[[i]])) == 0 &length(grep("% cover",st2835.txt[[i]])) == 0 &length(grep("average.{,15}cover",st2835.txt[[i]])) == 0&length(grep("species cover",st2835.txt[[i]])) == 0&length(grep("greater.{,15}cover",st2835.txt[[i]])) == 0&length(grep("lower.{,15}cover",st2835.txt[[i]])) == 0&length(grep("higher.{,15}cover ",st2835.txt[[i]])) == 0&length(grep("less.{,15}cover",st2835.txt[[i]])) == 0&length(grep("more.{,15}cover",st2835.txt[[i]])) == 0&length(grep(" cover of",st2835.txt[[i]])) == 0&length(grep(" cover decreased",st2835.txt[[i]])) == 0&length(grep(" cover increased",st2835.txt[[i]])) == 0&length(grep("decreased.{,15}cover",st2835.txt[[i]])) == 0&length(grep("..creased.{,15}cover",st2835.txt[[i]])) == 0&length(grep("less.{,25}cover",st2835.txt[[i]])) == 0&length(grep("more.{,25}cover",st2835.txt[[i]])) == 0 &length(grep("total.{,25}cover",st2835.txt[[i]])) == 0 &length(grep("proportion of.{,15}cover",st2835.txt[[i]])) == 0 &length(grep("cover was higher",st2835.txt[[i]])) == 0&length(grep("cover was lower",st2835.txt[[i]])) == 0&length(grep("% of the surface was cover",st2835.txt[[i]])) == 0){masterlog[i,4]=0}
    if(length(grep(" cover plot",st2835.txt[[i]])) > 0){masterlog[i,4]=0}

    #####survival
    if(length(grep("surviv..",st2835.txt[[i]])) > 0 ){masterlog[i,5]=1}
    if(length(grep("surviv..",st2835.txt[[i]])) == 0 ){masterlog[i,5]=0}

    #####reproductive success
    if(length(grep("breeding pairs ..creased",st2835.txt[[i]])) > 0 |length(grep("fledg",st2835.txt[[i]])) > 0 |length(grep("brood",st2835.txt[[i]])) > 0 |length(grep("% of egg",st2835.txt[[i]])) > 0 |length(grep("nestlings produced",st2835.txt[[i]])) > 0 |length(grep("proportion of females with chicks",st2835.txt[[i]])) > 0 |length(grep(" success ",st2835.txt[[i]])) > 0 | length(grep("clutch",st2835.txt[[i]])) > 0 | length(grep("significantly fewer (.*?) nests",st2835.txt[[i]])) > 0| length(grep("significantly less (.*?) nests",st2835.txt[[i]])) > 0| length(grep("significantly more (.*?) nests",st2835.txt[[i]])) > 0| length(grep("more nests",st2835.txt[[i]])) > 0 | length(grep("less nests",st2835.txt[[i]])) > 0 | length(grep("fewer nests",st2835.txt[[i]])) > 0){masterlog[i,6]=1}
    if(length(grep("breeding pairs ..creased",st2835.txt[[i]])) == 0 &length(grep("fledg",st2835.txt[[i]])) == 0 &length(grep("brood",st2835.txt[[i]])) == 0 &length(grep("% of egg",st2835.txt[[i]])) == 0 &length(grep("nestlings produced",st2835.txt[[i]])) == 0 &length(grep("proportion of females with chicks",st2835.txt[[i]])) == 0 &length(grep(" success ",st2835.txt[[i]])) == 0 & length(grep("clutch",st2835.txt[[i]])) == 0& length(grep("significantly fewer (.*?) nests",st2835.txt[[i]])) == 0& length(grep("significantly less (.*?) nests",st2835.txt[[i]])) == 0& length(grep("significantly more (.*?) nests",st2835.txt[[i]])) == 0& length(grep("more nests",st2835.txt[[i]])) == 0 & length(grep("less nests",st2835.txt[[i]])) == 0 & length(grep("fewer nests",st2835.txt[[i]])) == 0){masterlog[i,6]=0}

    #####mortality
```

```
    if(length(grep("% predation",st2835.txt[[i]])) > 0 |length(grep("mortality",st2835.txt[[i]])) > 0
|length(grep("predation rate",st2835.txt[[i]])) > 0 ){masterlog[i,7]=1}
    if(length(grep("% predation",st2835.txt[[i]])) == 0 &length(grep("mortality",st2835.txt[[i]])) == 0
&length(grep("predation rate",st2835.txt[[i]])) == 0 ){masterlog[i,7]=0}

    ###Error
    if(length(grep("url was not found",st2835.txt[[i]])) > 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}
    if(length(grep("server error",st2835.txt[[i]])) > 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}
    if(length(grep("summarised by",st2835.txt[[i]])) > 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}
    if(length(grep("no summary",st2835.txt[[i]])) > 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}

    ###blank text
    if(length(grep("t",st2835.txt[[i]])) == 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}
    if(length(grep("laboratory",st2835.txt[[i]])) == 0 &length(grep("comparison",st2835.txt[[i]])) == 0
&length(grep("study",st2835.txt[[i]])) == 0 & length(grep("experiment",st2835.txt[[i]])) == 0 &
length(grep("trial",st2835.txt[[i]])) == 0 ){masterlog[i,]=rep(NA,NCOL(masterlog))}

    masterlog[,8] = z
  }

  ###### find synopsis name
  masterlog = masterlog[which(is.na(masterlog[,1])==FALSE),]
  synops = regmatches(st2835tree.html, gregexpr('(?<=alt=\").*?(?=\">)', st2835tree.html, perl=T))
  if(length(synops[[1]]) == 0) {synops2 = NA}
  if(length(synops[[1]]) == 1) {synops2 = synops[[1]]}
  if(length(synops[[1]]) > 1 ) {synops2 = synops[[1]][1:(length(synops[[1]])-1)]}
  if(is.null(nrow(masterlog))==TRUE){
    masterlog[9] = synops2[1]
  }

  ##### find intervention name
  interv = unlist(c(regmatches(st2835tree.html, gregexpr('(?<=actions/).*?(?=</a>)',
st2835tree.html, perl=T))))
  interv = gsub('\">','',interv)
  interv = gsub('\\d','',interv)

  if(length(interv) == 0) {interv2 = NA}
  if(length(interv) == 1) {interv2 = interv}
  if(length(interv) > 1 ) {interv2 = interv[1:(length(interv))]}
  if(is.null(nrow(masterlog))==TRUE){
    masterlog[10] = interv2[1]
  }

  if(is.null(nrow(masterlog))==FALSE){
    masterlog[,9] = synops2[1:nrow(masterlog)]
    masterlog[,10] = interv2[1:nrow(masterlog)]
    if(length(which(duplicated(masterlog)==TRUE))>0){
      masterlog = masterlog[-which(duplicated(masterlog)==TRUE),]
    }
  }

  if(NROW(masterlog)>1 & NCOL(masterlog)>1){
    for(o in 1:NROW(masterlog)){
      if(is.na(masterlog[o,1])==TRUE){masterlog[o,]=rep(NA,NCOL(masterlog))}
    }
  }
  if(NROW(masterlog)<2 | NCOL(masterlog)<2){
    if(is.na(masterlog[1])==TRUE){masterlog=rep(NA,length(masterlog))}
  }
 }
```

```
  if(length(masterlog)==0){ masterlog = c(rep(NA,7),z,rep(NA,2))}
  return(masterlog)
}

#####Check function in serial
samp=sample(1:1698,1)
masterloop(samp)

#### Run in parallel
y = c(1:1697,1699:nstud) #webpage 1698 causes an error and contains and old study not assigned
to a synopsis

ncores=8
c1 <- makeCluster(ncores) ### number of cores your machine can run this in parallel

###tell cluster what packages it needs too
clusterEvalQ(c1,c(library(RCurl),library(tm),library(XML)))

master <-  parSapplyLB(c1,y,masterloop)

stopCluster(c1)

##### output results
master1 = do.call(rbind,master)

str(master)
str(master1)
colnames(master1)=c("density","abundance","diversity/richness","cover","survival","reprosuccess","
mortality","pageIDt","synopsis","intervention")
head(master1)
```

5 | Poor availability of context-specific evidence hampers decision-making in conservation

This chapter was published as:

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020. Poor availability of context-specific evidence hampers decision-making in conservation. Biol. Conserv. 248, 108666. https://doi.org/10.1016/j.biocon.2020.108666

Abstract

Evidence-based conservation relies on reliable and relevant evidence. Practitioners often prefer locally relevant studies whose results are more likely to be transferable to the context of planned conservation interventions. To quantify the availability of relevant evidence for amphibian and bird conservation we reviewed Conservation Evidence, a database of quantitative tests of conservation interventions. Studies were geographically clustered, and few locally conducted studies were found in Western sub-Saharan Africa, Russia, South East Asia, and Eastern South America. Globally there were extremely low densities of studies per intervention – fewer than one study within 2,000km of a given location. The availability of relevant evidence was extremely low when we restricted studies to those studying biomes or taxonomic orders containing high percentages of threatened species, compared to the most frequently studied biomes and taxonomic orders. Further constraining the evidence by study design showed that only 17-20% of amphibian and bird studies used reliable designs. Our results highlight the paucity of evidence on the effectiveness of conservation interventions, as well as the disparity between the availability of evidence for local contexts that are frequently studied versus those where conservation needs are greatest. Addressing the serious global shortfall in context-specific evidence requires a step change in the frequency of testing conservation interventions, greater use of reliable study designs and standardised metrics, and methodological advances to analyse patchy evidence bases.

Introduction

Tackling the biodiversity crisis with limited resources requires efficient and effective conservation action (Dirzo et al., 2014; Sutherland et al., 2004). To inform which conservation actions ('interventions') are effective and which are not, we need a large and reliable evidence base, ideally including large numbers of studies (replication of evidence; Fig.1A) with high internal validity (quality; Fig.1A) and external validity (relevance; Fig.1A). However, the limited resources available for conservation research mean that the evidence base for conservation is geographically and taxonomically biased (Christie et al., 2020; Donaldson et al., 2016; Murray et al., 2015; Spooner et al., 2015). This is likely to limit the quality and relevance of evidence and hinder effective decision-making (Cook et al., 2013b). Quantifying the availability of relevant, reliable studies is necessary to understand the strength of evidence upon which decisions are made, and to prioritise research on the effectiveness of conservation interventions.

The replication of evidence – the number of studies in the evidence base – is important as greater numbers of studies demonstrating repeatable and reproducible effectiveness will give us greater confidence in the overall strength of the evidence. Decision-makers should rightly be wary of basing decisions on a low number of studies where reproducible effectiveness has not been or cannot be demonstrated – particularly given the current reproducibility crisis (Begley and Ioannidis, 2015; Nosek and Errington, 2017; Open Science Collaboration, 2015). However, the overall number of studies is not the only indicator of the strength of the evidence, since studies with low internal validity (e.g., poor study designs) and/or external validity (i.e., low relevance) may not constitute reliable evidence.

The reliability of an evidence base – the internal validity of its studies – ultimately determines the overall quality of the evidence base and depends to a large extent on study design (Christie et al., 2019; de Palma et al., 2018; Spake and Doncaster, 2017). As the evidence base for conservation contains a wide variety of study designs (de Palma et al., 2018), there is likely to be variation in the reliability of inferences that can be drawn (Christie et al., 2019). This variation may lead scientists to make misleading recommendations to practitioners, ultimately reducing the effectiveness of conservation practice, and making it difficult for decision-makers to weigh the strength of evidence provided by different studies.

Practitioners and policymakers typically prefer to base their decisions on evidence that is relevant to their local context (i.e., with high external validity; Fig.1; Addison et al., 2016; Geijzendorffer et al., 2017; Gutzat and Dormann, 2020). For example, evidence that is drawn from a similar habitat, species, and socioeconomic context (to the one that a decision-maker is interested in) to maximise the likelihood that the findings of this evidence will apply there.

As a result, decision-makers typically use their own 'local knowledge' (e.g., Local Ecological Knowledge (LEK) or tacit knowledge), based on the experience or intuition of practitioners, stakeholders, and decision-makers to make decisions (Tanner et al., 2020; Wheeler and Root-Bernstein, 2020). To inform practitioners with locally valid evidence from the scientific literature to build upon this local knowledge and strengthen evidence-based decision-making, we need to consider the different factors that determine the relevance of scientific evidence to practitioners. The relevance of conservation studies to a given context will span multiple dimensions, including: (i) bioclimatic (i.e., similarity between habitats or regions); (ii) taxonomic/functional (i.e., similarity between taxa in terms of ecological function or taxonomic groups); and (iii) which metric was used to quantify the effectiveness of an intervention (i.e., the response variables or metrics of interest; Fig.1B). Other dimensions may also be important, such as the similarity between a study's and a practitioner's socioeconomic and political contexts, but we focus on the three dimensions depicted (Fig.1).

The first of these dimensions – bioclimatic relevance – refers to the similarity between the study ecosystem and the practitioner's ecosystem (Fig.1B). The second dimension – taxonomic/functional relevance – concerns the similarity between the focal taxa of a study and the taxa of interest to the practitioner (Fig.1B). Together, these determine the ecological similarity between study and practitioner local contexts. This is vital because responses to interventions will vary between ecosystems and taxa. For example, the effectiveness of artificial nest boxes varies between different countries and habitats (Finch et al., 2019), whilst the effectiveness of translocation for New Zealand robins (*Petroica australis*) is unlikely to be relevant to a practitioner translocating Kakapo (*Strigops habroptila*) as the latter is a flightless bird. Practitioners who are interested in broader functional groups (e.g., seed dispersers or pollinators), taxa (e.g., birds, amphibians), or even whole ecosystems, may focus more on the functional relevance rather than taxonomic similarity of studied species.

The third dimension of relevance is the metric used to measure the effectiveness of an intervention. Practitioners may be interested in different responses to interventions depending on their focus (e.g., species or ecosystem-level responses) and effectiveness may vary depending on the metric used (Capmourteres and Anand, 2016; Marshall et al., 2019). For example, at the ecosystem-level, the effectiveness of bird boxes may be measured using the species richness or diversity of birds using them (Caine and Marion, 1991), whilst at the species-level, the number of individuals (Brawn and Balda, 1988), fledglings (Male et al., 2006; Purcell et al., 1997), or brood size (Browne, 2006) may be measured. Similarly, the effectiveness of road mitigation interventions (e.g., tunnels or bridges) may be measured by the numbers of individuals of different species using the structures but could also be measured in terms of levels of road mortality (Helldin and Petrovan, 2019). Therefore, the type of metric

used by studies to measure effectiveness can have a major influence on the relevance of evidence.

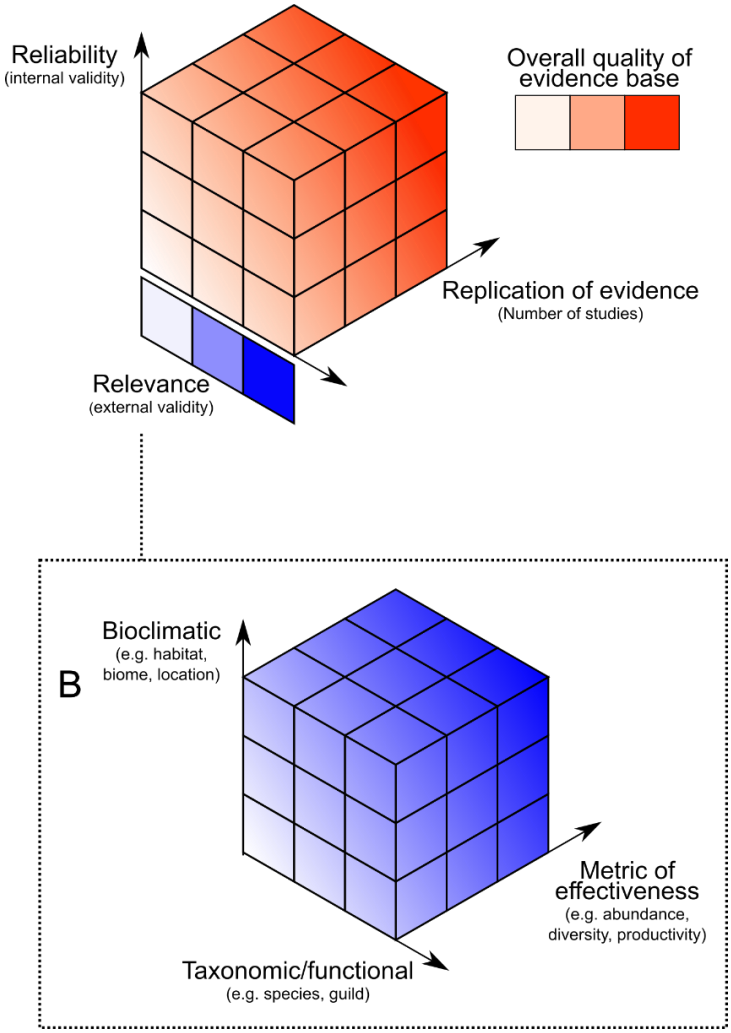


Figure 1 *– Framework of the desirable aspects of an ideal evidence base (stronger colours = more desirable). Fig.1A shows the three major desirable factors that an evidence base should have; large replication of evidence that is highly reliable (high internal validity) and highly relevant (high external validity). Fig.1B refers to the three dimensions that we will focus on that influence the overall relevance of evidence: i) bioclimatic (e.g., the study system), ii) taxonomic/functional (the study taxa) and iii) effectiveness measure (how you define and measure conservation success).*

Currently, we have a poor quantitative understanding of the availability of relevant and reliable studies in the literature that tests conservation interventions. In this study, we assess whether studies testing conservation interventions are distributed across different contexts (bioclimatically, taxonomically, and by the metric used to measure effectiveness) in ways that

reflect the needs of conservation (i.e., is research effort focused on testing interventions on threatened species or in locations where more threatened species occur?). We also quantify other desirable aspects of the evidence base for conservation in terms of the quantity and quality of locally relevant studies (i.e., how many studies test conservation actions within the locality of a given practitioner, and how many of these use reliable study designs?).

Materials and methods

Conservation Evidence database

We assessed the availability of relevant evidence for conservation practice using Conservation Evidence, a database of 5,525 publications as of January 2020 (Conservation Evidence, 2020a) that have quantitatively assessed the effectiveness of conservation interventions. Interventions are defined as management actions that a practitioner may undertake to benefit biodiversity (see Sutherland et al. (2019) for detailed methods). When we refer to the number of studies per intervention, we refer to the number of different tests of interventions – single publications may report multiple tests of different interventions. We assessed the availability of evidence for amphibians and birds based on synopses compiled in 2014 (n=419 studies; Smith and Sutherland, 2014) and 2012 (n=1,232 studies; Williams et al., 2013), respectively. More recent publications will obviously have increased the evidence base, but the broad patterns we quantify are unlikely to have changed in the intervening years. We excluded meta-analyses and reviews from our analyses as these typically cannot be attributed to a particular local context (e.g., biome or taxon). We also only included interventions for which studies were present in the database. Since 32% (n=33) of interventions for amphibians and 25% (n=80) of interventions for birds had no associated studies in the database (i.e., were untested or tests were unpublished) or only included reviews or meta-analyses, the following analyses are likely to be an optimistic assessment of the availability of evidence in conservation. We used R statistical software version 3.5.1 (R Core Team, 2019) for all analyses.

Local availability of studies by geographical distance

To calculate the average number of studies within a certain distance of somewhere a practitioner may wish to implement an intervention, we generated and then measured the distance of studies to 1,000 regularly spaced coordinates across the world. We regularly spaced coordinates over the terrestrial landmasses for birds, and within the combined extent of all amphibian species ranges for amphibians (IUCN, 2019). The spacing of coordinates was designed to represent the possible range of locations in which a practitioner might conduct an intervention to conserve amphibians or birds. Terrestrial landmasses were chosen for birds because although the combined distribution of all bird species is almost global, most

practitioners are likely to conduct interventions to conserve birds terrestrially. Although non-terrestrial interventions are carried out by practitioners, the vast area covered by the ocean would severely underestimate the availability of studies to a practitioner's likely location. 19 non-terrestrial interventions for birds were found in the database (e.g., 'use streamer lines to reduce seabird bycatch on longlines' or 'use high-visibility mesh on gillnets to reduce seabird bycatch') containing 33 studies in total – these were still included in our analysis as these studies tended to be conducted within close proximity to a terrestrial landmass (i.e., coastal). To account for coastal and island interventions for birds, we buffered the terrestrial landmasses used to regularly space coordinates by 1-degree (~111km depending on latitude; Fig.S9-S10). With the appropriate shapefiles for amphibians and birds, we first generated a regularly spaced grid of coordinates, checked which coordinates fell within the appropriate shapefile (from the IUCN (2019) for amphibians and OpenStreetMap (2019) for birds), and adjusted until we produced the desired number of regularly spaced coordinates (see Fig.S9-S12 for final maps of coordinates). We used R statistical software version 3.5.1 (R Core Team, 2019) and the packages sp (Bivand et al., 2013; Pebesma and Bivand, 2005), rgdal (Bivand et al., 2019) and rgeos (Bivand and Rundel, 2019) – R code to perform all analyses is available at https://doi.org/10.5281/zenodo.3634780.

We then calculated the Great Circle Distance from each study to each coordinate (this incorporates the curvature of the Earth when calculating distances) – we used the geosphere package (Hijmans, 2017) in R. Studies in each intervention were then binned into a series of categories based on the Great Circle Distance between studies and coordinates (100 km, 1,000 km and then every 1,000 km up to and including 19,000 km). We also calculated the 'Global Mean', which is the mean number of studies per intervention in the entire database – equivalent to approximately 20,000 km at the equator, the maximum distance separating any two coordinates. We then calculated the mean number of studies within each distance bin across all coordinates, as well as the number of studies that used different categories of study designs: i) any design, ii) Before-After (BA), Control-Impact (CI), Before-After Control-Impact (BACI) or Randomised Control-Impact (RCI); iii) CI, BACI or RCI; iv) BACI or RCI designs (see Materials and methods in Christie, Amano, Martin, Petrovan, et al. 2020 for definitions of each design).

We then repeated each analysis using the same number of coordinates (n=1,000), but this time by randomly selecting coordinates from amphibian and bird studies in the database (sampling with replacement from amphibian studies as there were fewer than 1,000). Using both approaches provided likely upper and lower bounds of evidence availability: regular coordinates likely underestimated the availability of evidence to practitioners, giving equal weighting to locations where conservation interventions are unlikely to occur (e.g., Antarctica)

and those that are more intensively managed (e.g., Europe). In contrast, using locations from existing publications will likely overestimate study availability as this assumes that practitioners only conduct interventions in locations where they have previously been tested.

We compared the results of the first analysis (regularly spaced coordinates) to the expected patterns we would observe if studies were regularly distributed. We did this by generating equal numbers of regularly spaced coordinates ('expected studies') as the number of amphibian and bird studies in total summed across interventions (564 and 1,560 coordinates, respectively, since some studies test multiple interventions) using the same methods and shapefiles as before. We then calculated the number of these 'expected studies' within each distance bin and divided by the total number of amphibian or bird interventions. This gave the expected mean number of studies per intervention in each distance bin had the studies been regularly spaced around the world.

To illustrate spatially explicit differences in the local availability of studies, we generated maps of the distance to the nearest study from each of the 1,000 regular coordinates for amphibians and birds. We used the longitude and latitude coordinates as centroid positions to display grid cells that were colour-coded by the distance to the nearest study (in km).

Context-specific availability of studies

To quantify the amount of relevant and reliable evidence on the effectiveness of different conservation interventions, we required metadata that described each study's local context and study design. By adapting previously described methods (Christie et al., 2020; Appendix S1), we extracted the biome, taxonomic order, and reported metric type used by each study (to quantify the number of relevant studies), as well as the broad category of study design used (to quantify the number of reliably designed studies). When metric metadata was extracted, we grouped similar metrics into the following nine metric types: count-based, diversity, activity-based, physiological, survival, reproductive success, education-based, regulation-based, and biomass (Appendix S1).

We quantified the number of studies per conservation intervention that met certain relevance and study design criteria, to give an estimate of the availability of relevant and reliable evidence. To ensure that we did not artificially constrain the number of studies per intervention for different subsets of studies (e.g., taxonomic order or biome), we grouped certain interventions that were focused on single taxa or habitats but were fundamentally the same type of intervention (e.g., 'create ponds for newts' and 'create ponds for toads' would be

grouped into ‘create ponds’; see Appendix S1 in Chapter 4 for these groupings). This resulted in a total of 71 and 226 interventions for amphibians and birds, respectively.

Using these interventions, we then undertook two analyses to quantify the availability of evidence under two different scenarios: (i) where we optimistically assume a given practitioner is interested in the most frequently studied local context; and (ii) where we assume that a given practitioner is interested in local contexts in which a greater percentage of species are threatened (i.e., those classified as Vulnerable, Endangered or Critically Endangered status on the IUCN (2019) Red List). We intersected shapefiles from the IUCN Red List with shapefiles of the world’s terrestrial biomes (Dinerstein et al., 2017) to determine the number of threatened species in each biome. We assumed that interventions could be tested by studies in any biome and on any taxonomic order – this will likely mean that our estimates for the second scenario are underestimates of study availability, for example, as certain interventions are unlikely to be conducted in certain biomes. However, we grouped interventions so that they were not defined as taxon or habitat-specific and used coarse criteria (biome and taxonomic order) to limit this underestimation.

The first analysis (Fig.S13) calculated the mean number of studies per intervention for both scenarios in terms of three separate relevance criteria: biome, taxonomic order, and metric. For the first scenario (i) we calculated the number of studies with the most frequently studied biome, order or metric relative to each intervention. For the second scenario (ii) to reflect conservation needs, we calculated the number of studies with a randomly selected biome, taxonomic order, or metric from a weighted list (averaged over 1,000 repeated runs). This weighted list was generated so that the probability of selection for biomes and taxonomic orders was determined by the number of threatened species that each biome and taxonomic order contained (i.e., those classified as Vulnerable, Endangered or Critically Endangered status on the IUCN (2019) Red List). The probability of selecting a given metric was relative to the number of times each metric was reportedly used in studies within each intervention.

For the second analysis (Fig.S14), we used a stepwise process to calculate the number of studies that met one or more of the relevance criteria – only carrying forward studies if they met all previous criteria. For example, considering the first scenario (most frequently studied context), we counted the number of studies featuring the most frequently studied biome, then studies featuring the most frequently studied biome AND taxonomic order, and then studies featuring the most frequently studied biome AND taxonomic order AND metric. We also repeated this for all possible orderings of biome, taxonomic order, and metric (Fig.5 and Figs.S2-S6), as well as for the second scenario (weighting towards biomes and taxonomic orders with greater percentages of threatened species). Taxonomic orders could only be

selected if at least one species in that order was present in the previously selected biome – we determined which taxonomic orders were present in each biome by intersecting shapefiles from the (IUCN, 2019) Red List with shapefiles of terrestrial biomes (Dinerstein et al., 2017). The same was true for biomes when taxonomic order was the first relevance criterion to be selected (i.e., only biomes in which that taxonomic order is present could be selected). In the final step, we also calculated the number of studies that used different categories of study designs (any design; BA, CI, BACI or RCI; CI, BACI or RCI; BACI or RCI). We chose to report the mean number of studies per intervention because using median values led to uninformative figures (as the majority of interventions had zero relevant studies for certain criteria) and did not facilitate our exploration of the data. We include figures showing median numbers (Fig.S1, S7 & S8) – our qualitative conclusions do not vary with the measure of central tendency used.

All data analysed in this study and code to repeat analyses are available from https://doi.org/10.5281/zenodo.3634779.

Results

We considered a total of 71 and 226 interventions for amphibians and birds (mean = 7.9 and 6.9 studies per intervention; Fig.2), respectively, that contained at least one study. Studies were not evenly distributed geographically; the mean number of amphibian and bird studies per intervention (large black circles in Fig.2) deviated, particularly for amphibians, from what we would have expected if the same number of studies were regularly distributed (orange triangles in Fig.2). On average, there was less than one study per intervention available within 2,000km from a given regular point (see vertical and horizontal lines on Fig.2). When restricting analyses to increasingly reliable designs, the availability of studies decreased substantially, with a higher proportion of amphibian studies using BA designs, compared to birds, but a smaller proportion using CI (see drop-offs from orange to blue, and blue to green lines, respectively; Fig.2).

When considering distance of studies to randomly selected study coordinates, the mean number of studies per intervention generally declined more gradually compared to a regular grid of coordinates (Fig.2), implying that studies are clustered in space. At distances below 5,000km these differences were particularly pronounced; for example, on average, 2.2 amphibian studies and 1.5 bird studies were within 2,000km of a random study coordinate, compared to only 0.5 amphibian studies and 0.2 bird studies within 2,000km of regularly spaced coordinate (see vertical and horizontal lines on Fig.2). This suggests that studies are slightly more clustered for amphibians than birds.

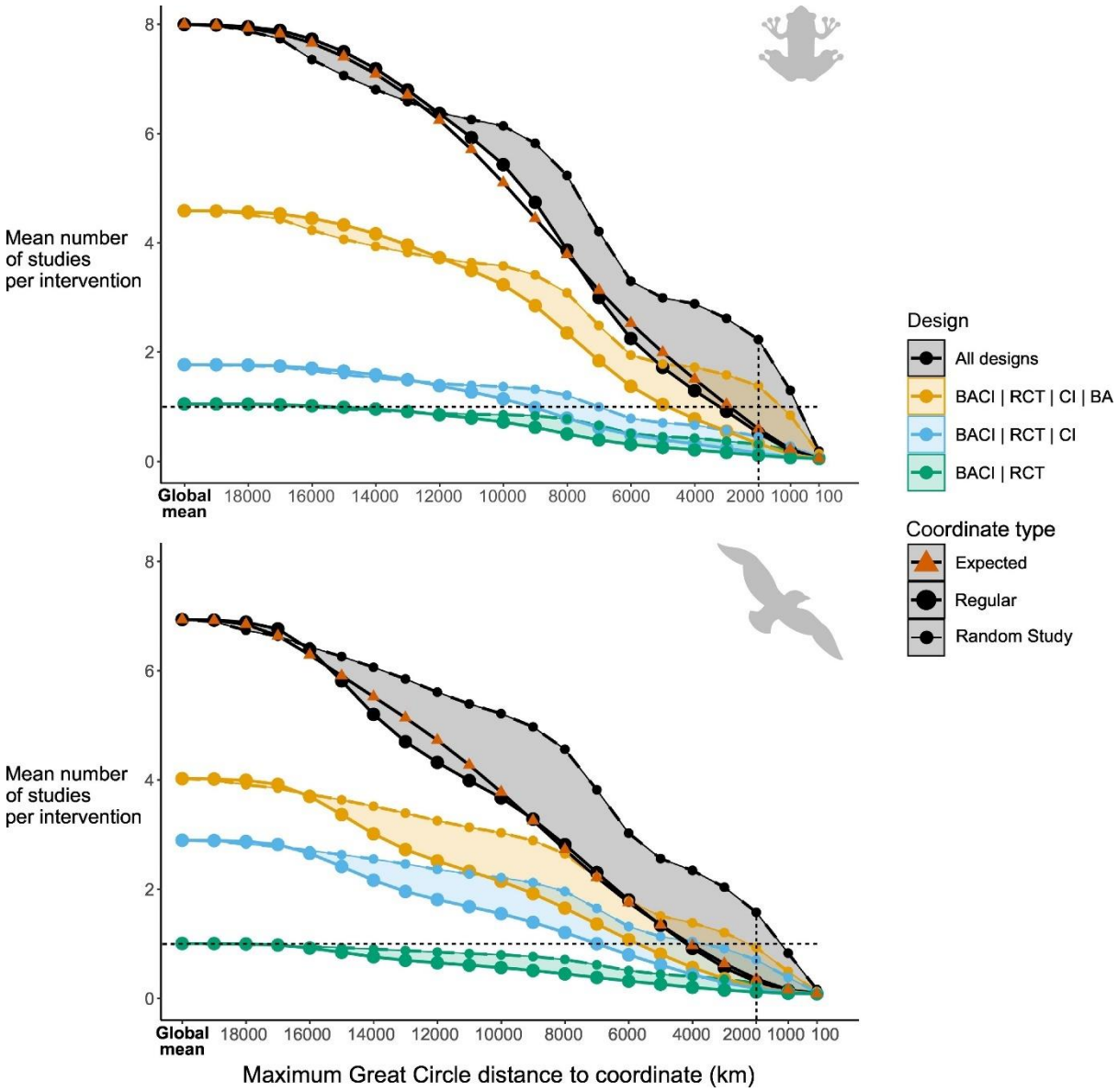


Figure 2 *– The mean number of amphibian and bird studies per intervention using different study designs found within a certain distance of different sets of coordinates. The maximum distance from a coordinate that a study must be conducted within is shown on the x axis, starting with the Global Mean (mean number of studies per intervention considering all studies in the database) and decreasing to a distance of 100 km. Regular coordinates (large circle, thick line) show the mean number of studies within a certain distance from a set of regularly distributed coordinates. Expected coordinates (orange triangle) mimic how the availability of studies would be expected to change if studies were regularly distributed (this is only shown for studies using any study design). Random Study coordinates (small circle, thin line) show the mean number of studies within a certain distance from a set of randomly selected coordinates where previous studies have been conducted. Dotted vertical and horizontal lines are placed to aid interpretation.*

Several regions in the combined range of all amphibian species contained few locally conducted studies; there were large distances (from 1500-4000km) to the nearest available study for regions including: Western sub-Saharan Africa, Central and North East South America, Russia, India, Sri Lanka, Bangladesh, and South East Asia (Fig.3). For birds, locations lacking locally conducted studies included: Western sub-Saharan Africa, Russia, Antarctica (except the Western Antarctic Peninsula), Eastern South America, and certain parts of South East Asia and Polynesia (Fig.3). For both amphibians and birds, most locations in North America, Europe, and Australasia had far smaller distances to the nearest study (<1500km, mostly less than 500km; Fig.3).

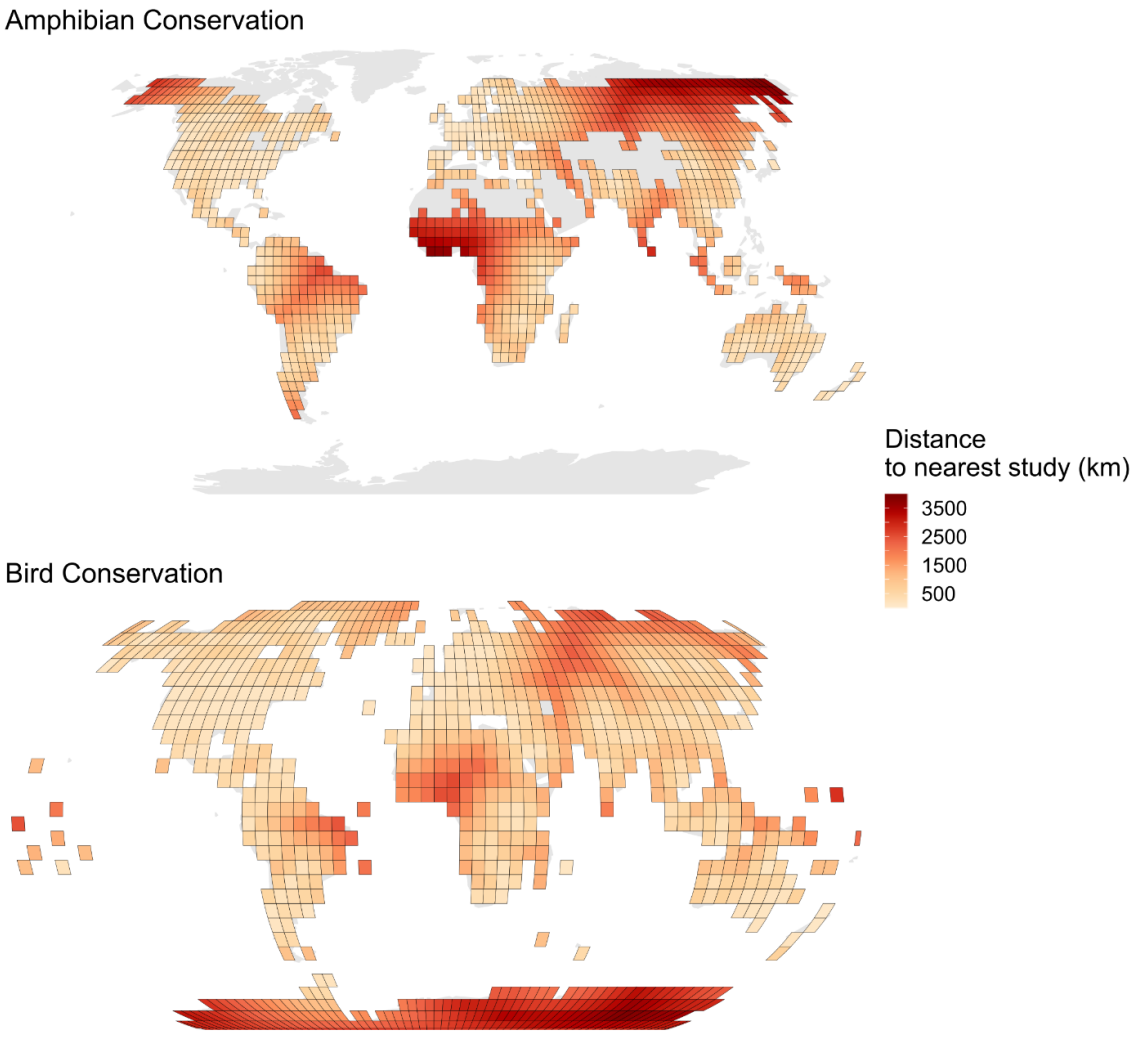


Figure 3 *– Maps illustrating the distance to the nearest amphibian or bird study in the conservation evidence database from 1,000 regularly spaced coordinates (at centroid position of grid cells) using a Robinson projection. Regularly spaced coordinates for amphibians sit within the combined extent of all extant amphibian species based on IUCN range maps (IUCN, 2019), whilst coordinates for birds sit within terrestrial land masses buffered by 1-degree to account for coastal interventions.*

The mean number of studies per intervention was substantially greater for the most frequently studied biome (Amphibians: 5.0 (95% Confidence Intervals (CIs) = [3.3, 6.8]; Birds: 3.5 [2.6, 4.5]), relative to each intervention, compared to biomes with higher percentages of species that are threatened (Amphibians: 0.4 [0.2, 0.6]; Birds: 0.4 [0.2, 0.5]; Fig.4). Similarly, the mean number of studies per intervention was substantially greater for the most frequently studied order in each intervention (Amphibians: 7.2 [4.8, 9.5]; Birds: 4.4 [3.2, 5.5]), compared to taxonomic orders with higher percentages of species that are threatened (Amphibians: 0.4 [0.0, 1.0]; Birds: 0.01 [0.0, 0.01]; Fig.4). There was a smaller difference in the mean number of studies per intervention between studies that used the most frequently used metric (Amphibians: 5.2 [3.7, 6.9]; Birds: 4.8 [3.4, 6.3]), relative to each intervention, and studies that used a randomly selected metric from within each intervention (Amphibians: 4.5 [3.4, 5.6]; Birds: 3.9 [2.9, 4.9]; Fig.4). The mean numbers of biomes, taxonomic orders, and metrics per intervention were 2.7 [2.2, 3.2], 2.6 [2.3, 3.0], and 3.1 [2.7, 3.5] for amphibians, respectively, and 2.4 [2.1, 2.7], 6.1 [5.1, 7.0], and 2.6 [2.3, 2.8] for birds, respectively.

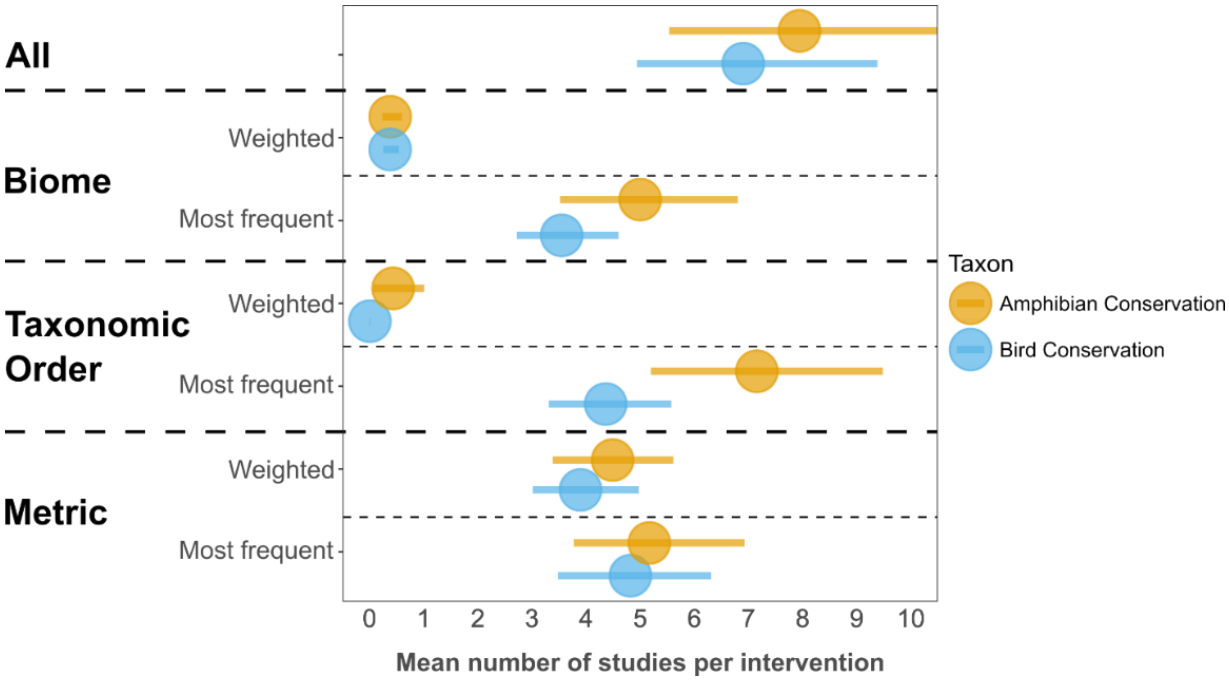


Figure 4 *– Mean number of studies per intervention when studies were counted based on whether they considered the most frequently studied biome, metric, or order, and whether they considered a randomly selected biome, metric, or taxonomic order from a weighted list. These weightings were based on the proportion of threatened species found in each biome or taxonomic order. 'All' indicates the mean number of studies per intervention when considering all studies. Error bars show bootstrapped 95% Confidence Intervals.*

The mean number of studies per intervention was also greater when we constrained by the most frequently studied biome, taxonomic order, and metric in a stepwise process Fig.5A),

compared to biomes and taxonomic orders with higher percentages of threatened species (Fig.5B). When we constrained by the most frequently studied biome, taxonomic order and metric, the greatest proportional decrease in the number of studies occurred once we further constrained by study design, by only counting studies using reliable BACI or RCI designs (on average, ~20% of amphibian studies and ~17% of bird studies that had met all previous criteria; Fig.5A). When we constrained by biomes and taxonomic orders with higher percentages of threatened species, the greatest proportional decreases occurred when constraining by taxonomic order, most notably for birds, and by biome (Fig.5B).

The sequence in which criteria were applied did not substantially affect the magnitude of the decrease in the number of studies – e.g., when biome was selected before or after taxonomic order and metric (Supplementary Information Fig.S2-S6). The overall decrease in studies from applying all relevance criteria (biome, taxonomic order, and metric) was similarly severe regardless of the sequence in which the criteria were applied (Supplementary Information Fig.S2-S6). For all sequences, constraining the evidence to studies that used reliable BACI or RCI designs reduced the mean number of studies to less than one study after constraining by the most frequently studied biome, taxonomic order, and metric (Fig.5A; Supplementary Information Fig.S2-S6). Doing the same after instead constraining by the biomes and taxonomic orders with higher percentages of threatened species reduced the mean number of studies to fewer than 0.01 studies with BACI or RCI designs (Fig.5B; Supplementary Information Fig.S2-S6).

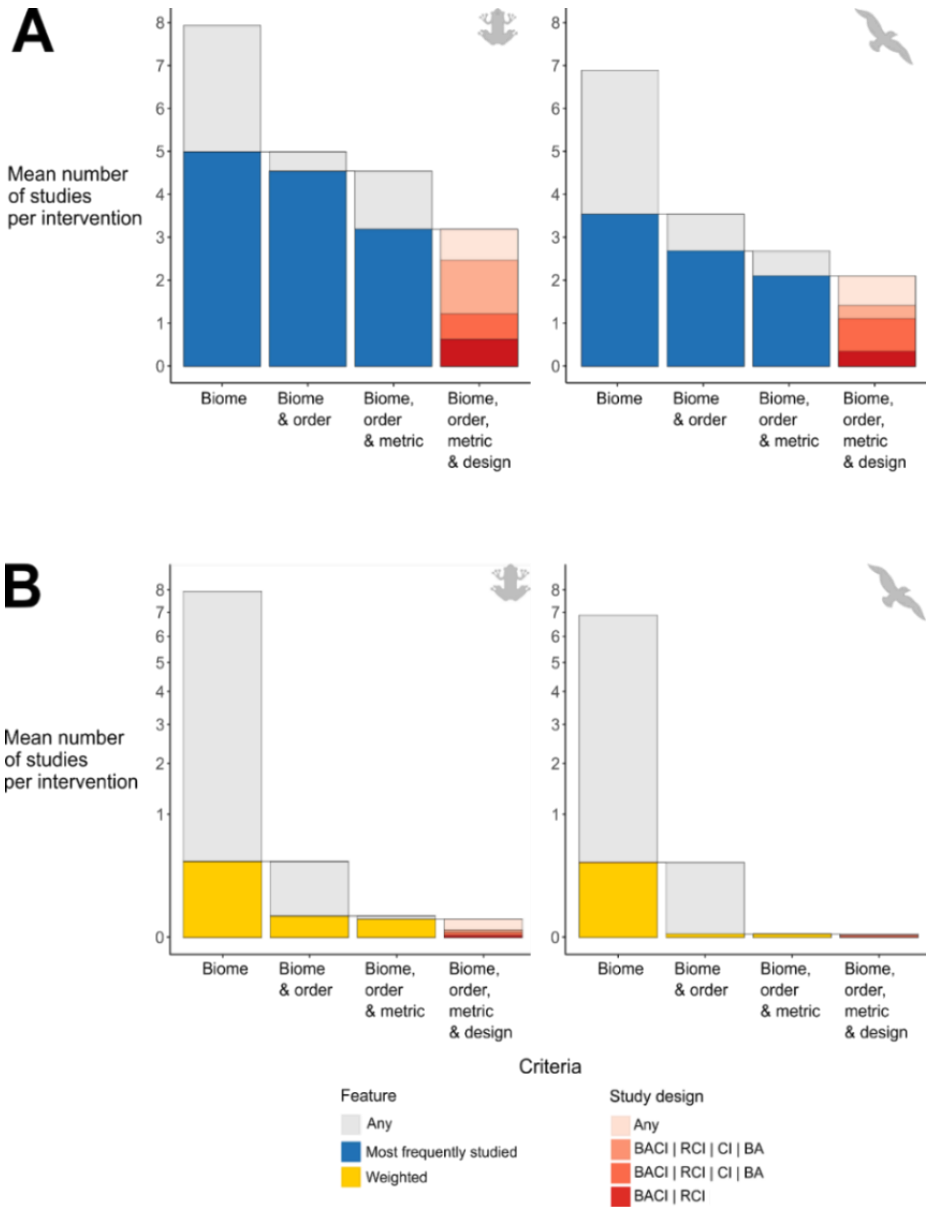


Figure 5 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied biome, taxonomic order, and metric relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies conducted in the most frequently studied biome were carried forward into the biome and order category. In panel B, studies with a selected biome, order, and metric were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used within each intervention. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs). Error bars were too small to be visible.*

Discussion

Our work demonstrates that not only is there a general paucity of studies testing conservation interventions, but that the distribution of these studies does not reflect conservation needs. Specifically, there is a lack of studies testing conservation interventions in biomes and for taxonomic orders containing high percentages of threatened amphibian and bird species. Given substantial declines of bird fauna (Rosenberg et al., 2019) and severe threats to amphibians (Grant et al., 2019), a better understanding of the effectiveness of interventions targeting threatened species is urgently required. Decision-makers are also likely to struggle to find locally conducted studies, let alone studies that use reliable study designs, particularly in Western sub-Saharan Africa, South East Asia, and Eastern South America. Addressing these deficits will be challenging, but there are several possible ways to improve the evidence base for conservation.

A fundamental problem that needs to be overcome in the long-term is the lack of studies testing conservation interventions. Williams et al. (2020) found that only 15% of studies from a representative sample of the conservation literature tested interventions. Evaluation of interventions should become common practice, both as a topic of academic research and as an activity for on-the-ground conservationists (Baylis et al., 2016). The publication of these tests, whether the results are positive, negative, or neutral, is critical to building a strong evidence base for conservation (Catalano et al., 2019). Current efforts to facilitate this include the Applied Ecology Resources repository (British Ecological Society, 2020), 'Evidence' articles in the journal Conservation Science and Practice (Society for Conservation Biology, 2020), and the journal Conservation Evidence (Conservation Evidence, 2020b).

Simply publishing more tests of conservation interventions, even at an increasing rate, is however, unlikely to solve the paucity of locally relevant studies. For example, even though adding 1,000 studies testing interventions on birds would increase the mean number of studies to approximately 11 studies across the current 226 interventions, these studies would still be spread thin across a myriad of local contexts where the need for conservation is often not the greatest (see also Wilson et al., 2016). Although Reboredo Segovia et al. (2020) suggest that the number of conservation science studies in tropical locations correlates with the number of threatened species, our results and earlier work (Christie et al., 2020) suggest this is not the case for conservation studies testing interventions. In fact, significantly fewer studies testing interventions were conducted in locations with greater numbers of threatened amphibian and bird species and there was a severe lack of studies from regions such as Africa, Russia, and South America (Christie et al., 2020). Several taxonomic orders of amphibians and birds were also found to be underrepresented, or even unrepresented, in the literature testing

conservation interventions relative to the percentage of threatened species they contain (e.g., caecilians and frogs, and parrots and songbirds; Christie et al., 2020). Therefore, we need concrete solutions enabling conservationists to generate and collate more experimental evidence on the effectiveness of conservation interventions for these underrepresented locations and taxa (Christie et al., 2020).

Funders, principal investigators, and heads of conservation organisations need to enhance and prioritise funding to test interventions in underrepresented regions identified by our study and previous work (Christie et al., 2020). Evidence synthesis also needs to incorporate more evidence from non-English language and grey literature publications to help address underrepresented local contexts (Amano et al., 2016; Amano and Sutherland, 2013) – for example, publications from over 317 non-English language journals are starting to be added to the Conservation Evidence database through the Transcending Language Barriers to Environmental Sciences project (translatE, 2020). This will help us understand whether the lack of locally conducted studies in underrepresented regions, such as South America and Russia (Fig.3), are due to language bias (e.g., most studies being published in Spanish, Portuguese or Russian rather than English), a genuine lack of testing of interventions, or a combination of both. Preliminary results suggest few studies testing conservation actions would be added from the non-English literature overall to change our major conclusions (Christie et al., 2020), but that they may help to address some geographic gaps in the English-language literature. Making concerted efforts to acquire grey literature from organisations and groups outside academia will also be important.

The low proportion of studies using reliable study designs, regardless of their relevance to a local context, is also challenging. That more reliably designed studies are concentrated in North America, Europe, and Australia compounds already severe taxonomic and biogeographical biases (Christie et al., 2020). If few reliably designed studies are available for informing conservation, decision-makers may have to consider a wider range of studies that may be less reliable or relevant, potentially reducing the effectiveness of decision-making and future practice (Slavin, 1995; Tugwell and Haynes, 2006; Whittaker, 2010). To increase the quality of studies available for decision-making, we must recognise that the quality of studies testing interventions may be limited in different ways. Studies evaluating mitigation efforts are often not constrained by cost, but rather by short timescales and their focus on meeting legislative requirements (for example, conserving legally protected species). Studies testing non-mitigation interventions will likely be more constrained by cost, as well as short timescales (e.g., PhD funding). Acknowledging how real-world constraints affect the choice of study design is essential to devising approaches to improving the evidence base for conservation. Whilst better training of early career scientists, consultants, and researchers in appropriate

study designs for causal inference may help, ultimately more regulatory and funder-led measures (e.g., requiring grantees to demonstrate rigorous study design) will be required (de Palma et al., 2018; Grant et al., 2019). Conservation interventions are too varied for strict guidelines or regulations to regulate the use of more reliable study designs, so we suggest that conservation researchers and practitioners think seriously about developing and following bespoke conservation-related or general scientific or clinical best-practice guidelines: for example, pre-registration and peer-review of methods (Parker et al., 2019). If practitioners are forced to rely on using less reliable study designs because of factors outside their control, study results must be reported with appropriate caution because their results may be biased by confounding factors and lead to misleading conclusions about the effectiveness of an intervention (Christie et al., 2019).

Given the general lack of evidence across conservation, there is also a need to use a standardised set of metrics to evaluate conservation effectiveness (McQuatters-Gollop et al., 2019). Using a diversity of metrics may be necessary to assess multiple important aspects of an intervention's effectiveness, but a lack of consistency in the metrics used to report results often makes the evidence base difficult to synthesise – especially if different metrics yield different results (Mace and Baillie, 2007). Prioritisation of the most relevant metrics of effectiveness for different interventions with input from decision-makers and practitioners is essential to facilitate inter-study comparisons (McQuatters-Gollop et al., 2019). Initiatives aiming to do this are underway in topics such as fishery habitats (Lederhouse and Link, 2016) and protected areas (Nolte and Agrawal, 2013; Pomeroy et al., 2004), and are supported by the Essential Biodiversity Variables framework (Jetz et al., 2019). Funders could help strengthen these efforts by requiring grantees to follow such initiatives and use consistent metrics when evaluating interventions. Preregistration of research plans could also provide the opportunity for the scientific community to direct researchers towards appropriate, consistent metrics to evaluate conservation interventions (Parker et al., 2019).

Increasing the size and quality of the evidence base for conservation decision-making will be a slow process, but conservation practitioners need to make decisions now. Until the evidence base improves, excluding studies from evidence syntheses because they do not meet certain quality or relevance criteria could lead to little or no evidence being used to inform conservation efforts (Davies and Gray, 2015; Gurevitch and Hedges, 1999; Lortie et al., 2015). Moreover, studies that do not meet these criteria may still provide useful evidence, particularly in the absence of more relevant and reliable studies (Burivalova et al., 2019; Cook et al., 2013a; Gough and White, 2018).

We need novel approaches to rigorously synthesising studies that vary considerably in their relevance and reliability to maximise the use of the current imperfect evidence base. We believe that weighting approaches in both quantitative meta-analyses and more qualitative evidence synthesis would help maximise the number of studies available, while giving greater influence to studies with desirable characteristics. This could involve giving greater influence to more reliably designed studies (e.g., using weights that incorporate study design bias and variance (Christie et al. 2020; Christie et al., 2019) and evidence hierarchies from Mupepele et al. (2016)), and giving more weight to more relevant studies (e.g., weighting by the relevance of studies to a decision-maker's local context, as proposed in healthcare by Kneale et al. (2019). These approaches are being pioneered at www.metadataset.com where users can perform interactive, dynamic meta-analyses (Shackelford et al., 2021) by defining the weights different studies receive based on their study design and relevance to the user's local context. To generate objective weights of study relevance that reflect the likely generalisability of study results, we need studies which help us to understand how generalisability varies between interventions for different ecological (e.g., artificial nest boxes; Finch et al., 2019), socioeconomic, and political contexts. Understanding why some interventions work in certain contexts and not others is fundamentally important for effective evidence-based decision-making (Grant et al., 2019).

Overall, we have shown that the literature testing conservation interventions does not reflect the needs of conservation (i.e., to prioritise the conservation of threatened species). The serious lack of locally relevant and reliable evidence on the effectiveness of different conservation interventions presents several major challenges to decision-making in conservation. We hope that the conservation community can work together to improve the state of the evidence base for conservation based on our recommendations, as this will require much greater collaboration between research and practice. Testing interventions needs to become more routine, use a more standardised suite of metrics and reliable study designs, and, most importantly, focus on the locations and taxa where evidence is most needed to inform conservation action. In the meantime, we need to explore ways to better analyse the current patchy evidence base of conservation and ensure that we can support the shift towards more evidence-based policy and practice at a local level.

References

Addison, P.F.E., Cook, C.N., de Bie, K., 2016. Conservation practitioners' perspectives on decision triggers for evidence-based management. Journal of Applied Ecology 53, 1351–1357. https://doi.org/10.1111/1365-2664.12734

Amano, T., González-Varo, J.P., Sutherland, W.J., 2016. Languages Are Still a Major Barrier to Global Science. PLOS Biology 14, e2000933. https://doi.org/10.1371/journal.pbio.2000933

Amano, T., Sutherland, W.J., 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. Proceedings of the Royal Society B: Biological Sciences 280, 20122649. https://doi.org/10.1098/rspb.2012.2649

Baylis, K., Honey-Rosés, J., Börner, J., Corbera, E., Ezzine-de-Blas, D., Ferraro, P.J., Lapeyre, R., Persson, U.M., Pfaff, A., Wunder, S., 2016. Mainstreaming Impact Evaluation in Nature Conservation. Conservation Letters 9, 58–64. https://doi.org/10.1111/conl.12180

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in Science. Circulation Research 116, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Bivand, R., Keitt, T., Rowlingson, B., 2019. rgdal: Bindings for the "Geospatial" Data Abstraction Library R package version 1.4-8.

Bivand, Roger.S., Rundel, C., 2019. rgeos: Interface to Geometry Engine - Open Source ('GEOS') R package version 0.5-2.

Bivand, R.S., Pebesma, E., Gomez-Rubio, V., 2013. Applied spatial data analysis with R, Second. ed. Springer, New York.

Brawn, J.D., Balda, R.P., 1988. Population biology of cavity nesters in northern Arizona: do nest sites limit breeding densities? The Condor 90, 61–71. https://doi.org/10.2307/1368434

British Ecological Society, 2020. Applied Ecological Resources repository and Ecological Solutions and Evidence journal [WWW Document]. URL https://besjournals.onlinelibrary.wiley.com/journal/26888319 (accessed 1.10.20).

Browne, S.J., 2006. Effect of nestbox construction and colour on the occupancy and breeding success of nesting tits Parus spp. Bird Study 53, 187–192. https://doi.org/10.1080/00063650609461432

Burivalova, Z., Miteva, D., Salafsky, N., Butler, R.A., Wilcove, D.S., 2019. Evidence Types and Trends in Tropical Forest Conservation Literature. Trends in Ecology and Evolution 34, 669–679. https://doi.org/10.1016/j.tree.2019.03.002

Caine, L.A., Marion, W.R., 1991. Artificial Addition of Snags and Nest Boxes to Slash Pine Plantations (Colocacion de maderos y cajas de anidamiento en plantaciones de Pinus elliottii). Journal of Field Ornithology 62, 97–106. https://www.jstor.org/stable/4513610

Capmourteres, V., Anand, M., 2016. “Conservation value”: a review of the concept and its quantification. Ecosphere 7, e01476. https://doi.org/10.1002/ecs2.1476

Catalano, A.S., Lyons-White, J., Mills, M.M., Knight, A.T., 2019. Learning from published project failures in conservation. Biological Conservation 238, 108223. https://doi.org/https://doi.org/10.1016/j.biocon.2019.108223

Christie, A.P., Abecasis, D., Adjeroud, M., Alonso, J.C., Amano, T., Anton, A., Baldigo, B.P., Barrientos, R., Bicknell, J.E., Buhl, D.A., Cebrian, J., Ceia, R.S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M.D., Davoult, D., De Backer, A., Donovan, M.K., Eddy, T.D., França, F.M., Gardner, J.P.A., Harris, B.P., Huusko, A., Jones, I.L., Kelaher, B.P., Kotiaho, J.S., López-Baucells, A., Major, H.L., Mäki-Petäys, A., Martín, B., Martín, C.A., Martin, P.A., Mateos-Molina, D., McConnaughey, R.A., Meroni, M., Meyer, C.F.J., Mills, K., Montefalcone, M., Noreika, N., Palacín, C., Pande, A., Pitcher, C.R., Ponce, C., Rinella, M., Rocha, R., Ruiz-Delgado, M.C., Schmitter-Soto, J.J., Shaffer, J.A., Sharma, S., Sher, A.A., Stagnol, D., Stanley, T.R., Stokesbury, K.D.E., Torres, A., Tully, O., Vehanen, T., Watts, C., Zhao, Q., Sutherland, W.J., 2020. Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. Nature Communications 11, 6377. https://doi.org/10.1038/s41467-020-20142-y

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R., Sutherland, W.J., 2020. The challenge of biased evidence in conservation. Conservation Biology cobi.13577. https://doi.org/10.1111/cobi.13577

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Conservation Evidence, 2020a. Conservation Evidence [WWW Document]. URL www.conservationevidence.com (accessed 2.4.20).

Conservation Evidence, 2020b. Conservation Evidence journal [WWW Document]. URL https://www.conservationevidence.com/collection/view

Cook, C.N., Mascia, M.B., Schwartz, M.W., Possingham, H.P., Fuller, R.A., 2013a. Achieving Conservation Science that Bridges the Knowledge–Action Boundary. Conservation Biology 27, 669–678. https://doi.org/10.1111/cobi.12050

Cook, C.N., Possingham, H.P., Fuller, R.A., 2013b. Contribution of Systematic Reviews to Management Decisions. Conservation Biology 27, 902–915. https://doi.org/10.1111/cobi.12114

Davies, G.M., Gray, A., 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). Ecology and Evolution 5, 5295–5304. https://doi.org/10.1002/ece3.1782

de Palma, A., Sanchez-Ortiz, K., Martin, P.A., Chadwick, A., Gilbert, G., Bates, A.E., Börger, L., Contu, S., Hill, S.L.L., Purvis, A., 2018. Challenges With Inferring How Land-Use Affects Terrestrial Biodiversity: Study Design, Time, Space and Synthesis. Next Generation Biomonitoring 58, 163–199. https://doi.org/https://doi.org/10.1016/bs.aecr.2017.12.004

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.P.B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F., Saleem, M., 2017. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. BioScience 67, 534–545. https://doi.org/10.1093/biosci/bix014

Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J.B., Collen, B., 2014. Defaunation in the Anthropocene. Science 345, 401–406. https://doi.org/10.1126/science.1251817

Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J., Kerr, J.T., 2016. Taxonomic bias and international biodiversity conservation research. FACETS 1, 105–113. https://doi.org/10.1139/facets-2016-0011

Finch, T., Branston, C., Clewlow, H., Dunning, J., Franco, A.M.A., Račinskis, E., Schwartz, T., Butler, S.J., 2019. Context-dependent conservation of the cavity-nesting European Roller. Ibis 161, 573–589. https://doi.org/10.1111/ibi.12650

Geijzendorffer, I.R., van Teeffelen, A.J.A., Allison, H., Braun, D., Horgan, K., Iturrate-Garcia, M., Santos, M.J., Pellissier, L., Prieur-Richard, A.-H., Quatrini, S., 2017. How can global conventions for biodiversity and ecosystem services guide local conservation actions? Current Opinion in Environmental Sustainability 29, 145–150. https://doi.org/10.1016/j.cosust.2017.12.011

Gough, D., White, H., 2018. Evidence Standards and Evidence Claims in Web Based Research Portals. Centre for Homelessness Impact. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/CFHI_EVIDENCE_STANDARDS_REPORT_V15_PRINT.pdf

Grant, E.H.C., Muths, E., Schmidt, B.R., Petrovan, S.O., 2019. Amphibian conservation in the Anthropocene. Biological Conservation 236, 543–547. https://doi.org/https://doi.org/10.1016/j.biocon.2019.03.003

Gurevitch, J., Hedges, L. v., 1999. Statistical Issues in Ecological Meta-analyses. Ecology 80, 1142–1149. https://doi.org/10.1890/0012-9658(1999)080[1142:SIIEMA]2.0.CO;2

Gutzat, F., Dormann, C.F., 2020. Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice. Environmental Management 66, 435–449. https://doi.org/10.1007/s00267-020-01312-6

Helldin, J.O., Petrovan, S.O., 2019. Effectiveness of small road tunnels and fences in reducing amphibian roadkill and barrier effects at retrofitted roads in Sweden. PeerJ 7, e7518. https://doi.org/10.7717/peerj.7518

Hijmans, R.J., 2017. geosphere: Spherical Trigonometry. R package version 1.5-7.

IUCN, 2019. IUCN Red List [WWW Document]. URL https://www.iucnredlist.org/ (accessed 11.12.19).

Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller, G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S., Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. Nature Ecology and Evolution 3, 539–551. https://doi.org/10.1038/s41559-019-0826-1

Kneale, D., Thomas, J., O'Mara-Eves, A., Wiggins, R., 2019. How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision making? Research Synthesis Methods 10, 44–56. https://doi.org/10.1002/jrsm.1320

Lederhouse, T., Link, J.S., 2016. A Proposal for Fishery Habitat Conservation Decision-Support Indicators. Coastal Management 44, 209–222. https://doi.org/10.1080/08920753.2016.1163176

Lortie, C.J., Stewart, G., Rothstein, H., Lau, J., 2015. How to critically read ecological meta-analyses. Research Synthesis Methods 6, 124–133. https://doi.org/10.1002/jrsm.1109

Mace, G.M., Baillie, J.E.M., 2007. The 2010 biodiversity indicators: Challenges for science and policy. Conservation Biology 21, 1406–1413. https://doi.org/10.1111/j.1523-1739.2007.00830.x

Male, S.K., Jones, J., Robertson, R.J., 2006. Effects of nest-box density on the behavior of Tree Swallows during nest building. Journal of Field Ornithology 77, 61–66. https://doi.org/10.1111/j.1557-9263.2006.00006.x

Marshall, E., Wintle, B.A., Southwell, D., Kujala, H., 2019. What are we measuring? A review of metrics used to describe biodiversity in offsets exchanges. Biological Conservation 108250. https://doi.org/https://doi.org/10.1016/j.biocon.2019.108250

McQuatters-Gollop, A., Mitchell, I., Vina-Herbon, C., Bedford, J., Addison, P.F.E., Lynam, C.P., Geetha, P.N., Vermeulan, E.A., Smit, K., Bayley, D.T.I., Morris-Webb, E., Niner, H.J., Otto, S.A., 2019. From Science to Evidence – How Biodiversity Indicators Can Be Used for Effective Marine Conservation Policy and Management. Frontiers in Marine Science 6, 1–16. https://doi.org/10.3389/fmars.2019.00109

Mupepele, A.-C., Walsh, J.C., Sutherland, W.J., Dormann, C.F., 2016. An evidence assessment tool for ecosystem services and conservation studies. Ecological Applications 26, 1295–1301. https://doi.org/10.1890/15-0595

Murray, H.J., Green, E.J., Williams, D.R., Burfield, I.J., de Brooke, M.L., 2015. Is research effort associated with the conservation status of European bird species? Endangered Species Research 27, 193–206. https://doi.org/10.3354/esr00656

Nolte, C., Agrawal, A., 2013. Linking Management Effectiveness Indicators to Observed Effects of Protected Areas on Fire Occurrence in the Amazon Rainforest. Conservation Biology 27, 155–165. https://doi.org/10.1111/j.1523-1739.2012.01930.x

Nosek, B.A., Errington, T.M., 2017. Making sense of replications. eLife 6, e23383. https://doi.org/10.7554/eLife.23383

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716. https://doi.org/10.1126/science.aac4716

OpenStreetMap [WWW Document], 2019. URL http://openstreetmapdata.com/data/land-polygons (accessed 12.14.19).

Parker, T., Fraser, H., Nakagawa, S., 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology 33, 747–750. https://doi.org/https://doi.org/10.1111/cobi.13342

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5.

Pomeroy, R.S., Parks, J.E., Watson, L.M., 2004. How is your MPA doing?: a guidebook of natural and social indicators for evaluating marine protected area management effectiveness. IUCN, Gland.

Purcell, K.L., Verner, J., Lewis W, O., 1997. A comparison of the breeding ecology of birds nesting in boxes and tree cavities. The Auk 114, 646–656.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Reboredo Segovia, A.L., Romano, D., Armsworth, P.R., 2020. Who studies where? Boosting tropical conservation research where it is most needed. Frontiers in Ecology and the Environment fee.2146. https://doi.org/10.1002/fee.2146

Rosenberg, K. v, Dokter, A.M., Blancher, P.J., Sauer, J.R., Smith, A.C., Smith, P.A., Stanton, J.C., Panjabi, A., Helft, L., Parr, M., Marra, P.P., 2019. Decline of the North American avifauna. Science eaaw1313. https://doi.org/10.1126/science.aaw1313

Shackelford, G.E., Martin, P.A., Hood, A.S.C., Christie, A.P., Kulinskaya, E., Sutherland, W.J., 2021. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biology 19, 33. https://doi.org/10.1186/s12915-021-00974-w

Slavin, R.E., 1995. Best evidence synthesis: An intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48, 9–18. https://doi.org/10.1016/0895-4356(94)00097-A

Smith, R.K., Sutherland, W.J., 2014. Amphibian conservation: global evidence for the effects of interventions. Pelagic Publishing Ltd., Exeter.

Society for Conservation Biology, 2020. Conservation Science and Practice [WWW Document]. URL https://conbio.onlinelibrary.wiley.com/journal/25784854 (accessed 1.10.20).

Spake, R., Doncaster, C.P., 2017. Use of meta-analysis in forest biodiversity research: key challenges and considerations. Forest Ecology and Management 400, 429–437. https://doi.org/10.1016/j.foreco.2017.05.059

Spooner, F., Smith, R.K., Sutherland, W.J., 2015. Trends, biases and effectiveness in reported conservation interventions. Conservation Evidence 12, 2–7. http://mobile.www.conservationevidence.com/reference/download/5494

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19, 305–308. https://doi.org/https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

translateE, 2020. translateE - Transcending Language Barriers to Environmental Sciences [WWW Document]. URL https://researchers.uq.edu.au/research-project/35572 (accessed 1.11.20).

Tanner, L., Mahajan, S.L., Becker, H., DeMello, N., Komuhangi, C., Mills, M., Masuda, Y., Wilkie, D., Glew, L., 2020. Making better decisions: How to use evidence in a complex world. The Research People and the Alliance for Conservation Evidence and Sustainability. https://www.allianceconservationevidence.org/s/Making_better_decisions_ACES.pdf

Tugwell, P., Haynes, R.B., 2006. Assessing claims of causation, in: Tugwell, B., Haynes, R.B., Sackett, D.L., Guyatt, G.H., Tugwell, P. (Eds.), Clinical epidemiology: how to do clinical practice research. The University of Chicago Press Philadelphia, Pennsylvania, pp. 356–387.

Wheeler, H.C., Root-Bernstein, M., 2020. Informing decision-making with Indigenous and local knowledge and science. Journal of Applied Ecology 57, 1634–1643. https://doi.org/10.1111/1365-2664.13734

Whittaker, R.J., 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91, 2522–2533. https://doi.org/10.1890/08-0968.1

Williams, D.R., Balmford, A., Wilcove, D.S., 2020. The past and future role of conservation science in saving biodiversity. Conservation Letters 13, 1–7. https://doi.org/10.1111/conl.12720

Williams, D.R., Pople, R.G., Showler, D.A., Dicks, L. v, Child, M.F., Zu Ermgassen, E.K.H.J., Sutherland, W.J., 2013. Bird Conservation: Global evidence for the effects of interventions. Pelagic Publishing Ltd., Exeter.

Wilson, K.A., Auerbach, N.A., Sam, K., Magini, A.G., Moss, A.St.L., Langhans, S.D., Budiharta, S., Terzano, D., Meijaard, E., 2016. Conservation Research Is Not Happening Where It Is Most Needed. PLOS Biology 14, e1002413. https://doi.org/10.1371/journal.pbio.1002413

Supplementary Information

Figure S1

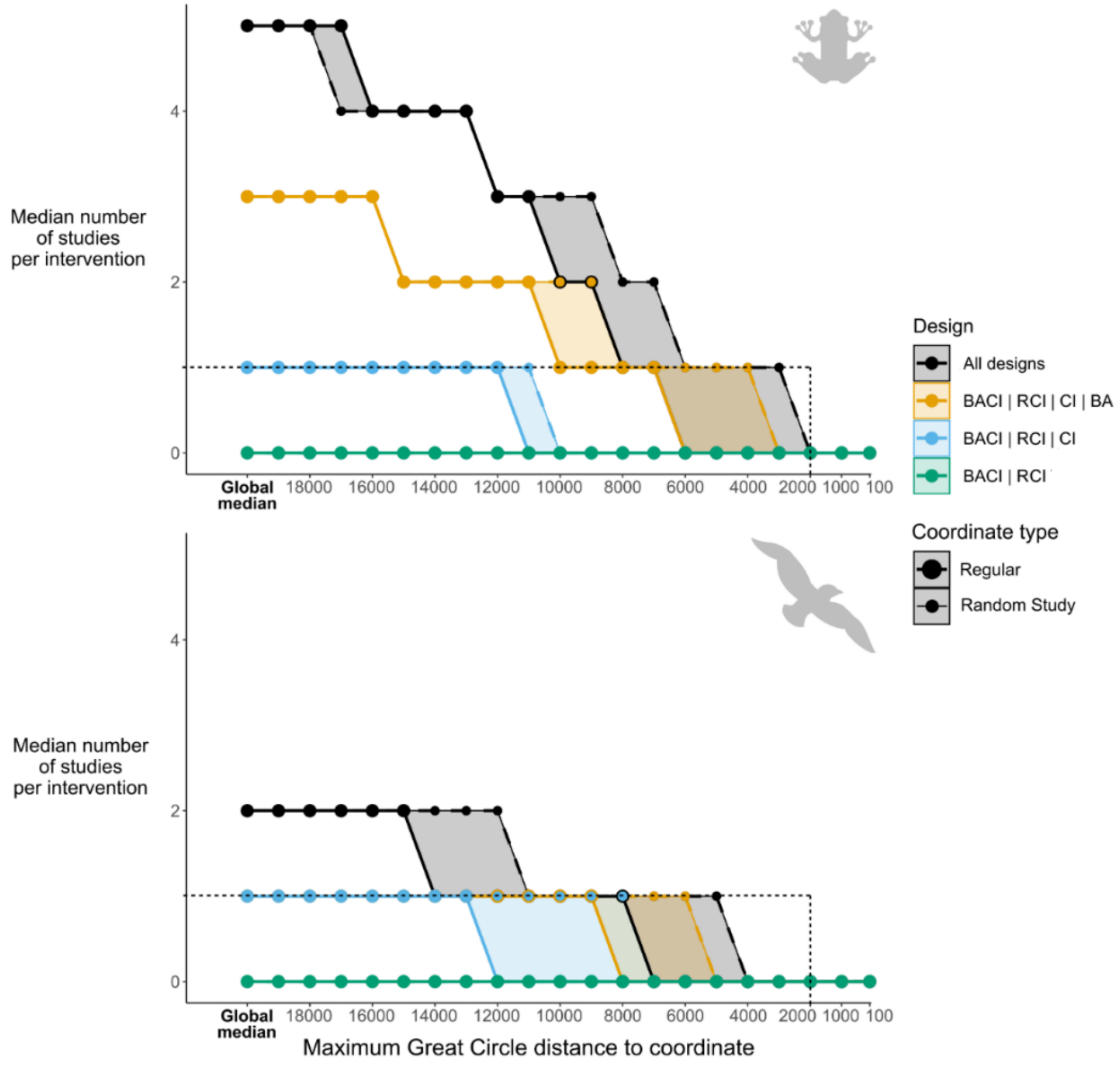


Figure S1 *– The median number of amphibian and bird studies per intervention using different study designs found within a certain distance of different sets of coordinates. The maximum distance that a study can be is shown on the x axis, starting with the Global Median (median number of studies per intervention considering all studies in the database) and decreasing to a distance of 100 km. Regular coordinates (large circle, thick line) show the median number of studies within a certain distance from a set of regularly distributed coordinates. Random Study coordinates (small circle, thin line) show the median number of studies within a certain distance from a set of randomly selected coordinates where previous studies have been conducted. Dotted vertical and horizontal lines are placed to aid interpretation.*

Figures S2-S6

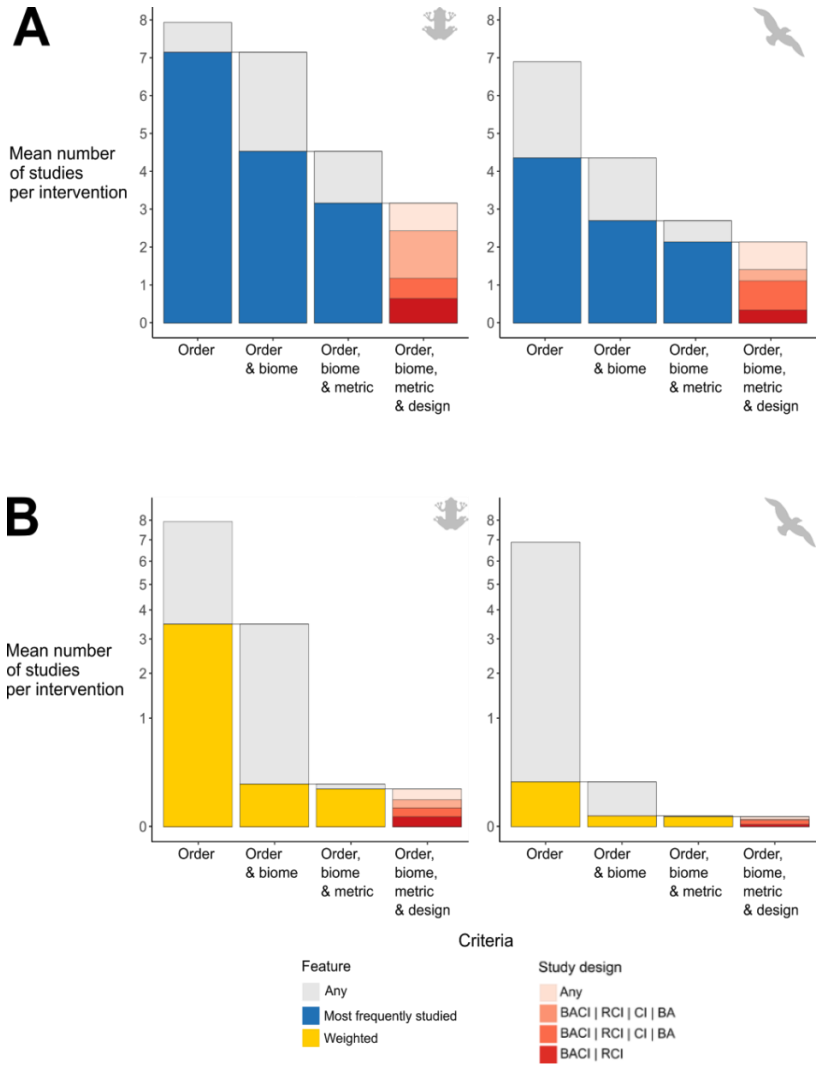


Figure S2 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied taxonomic order, biome and metric relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies with the most frequently studied order were carried forward into the order and biome category. In panel B, studies with a selected taxonomic order, biome and metric were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

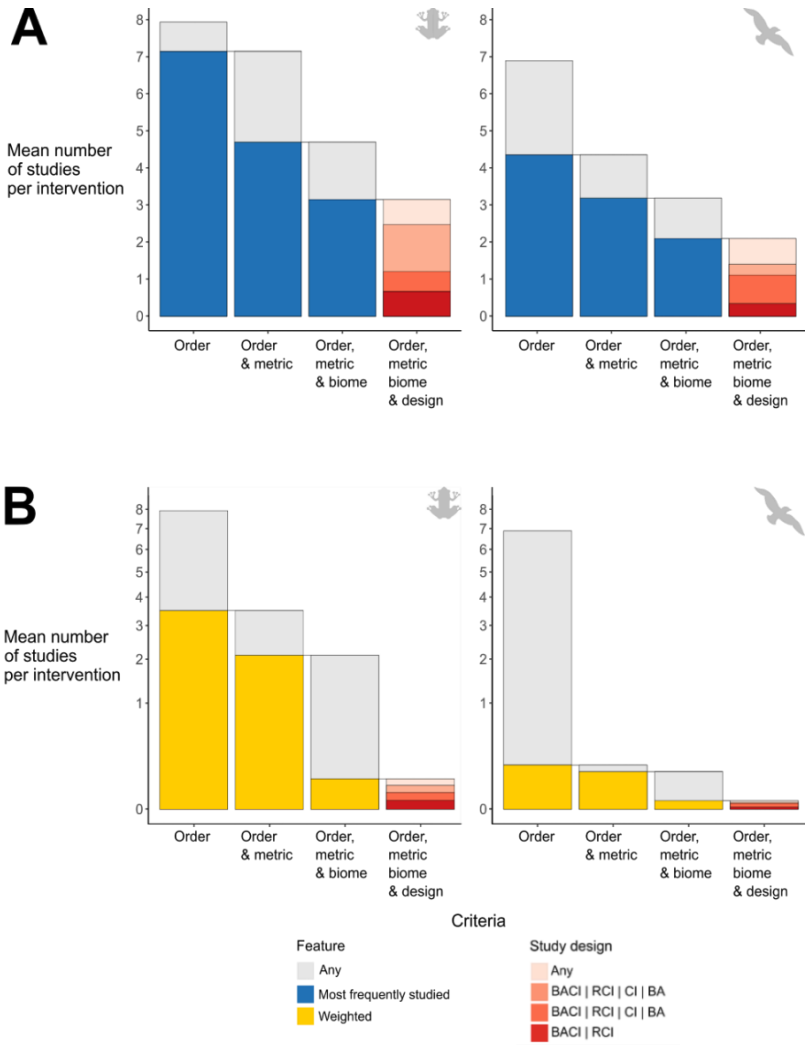


Figure S3 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied taxonomic order, metric, and biome relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies with the most frequently studied order were carried forward into the order and metric category. In panel B, studies with a selected taxonomic order, metric and biome were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

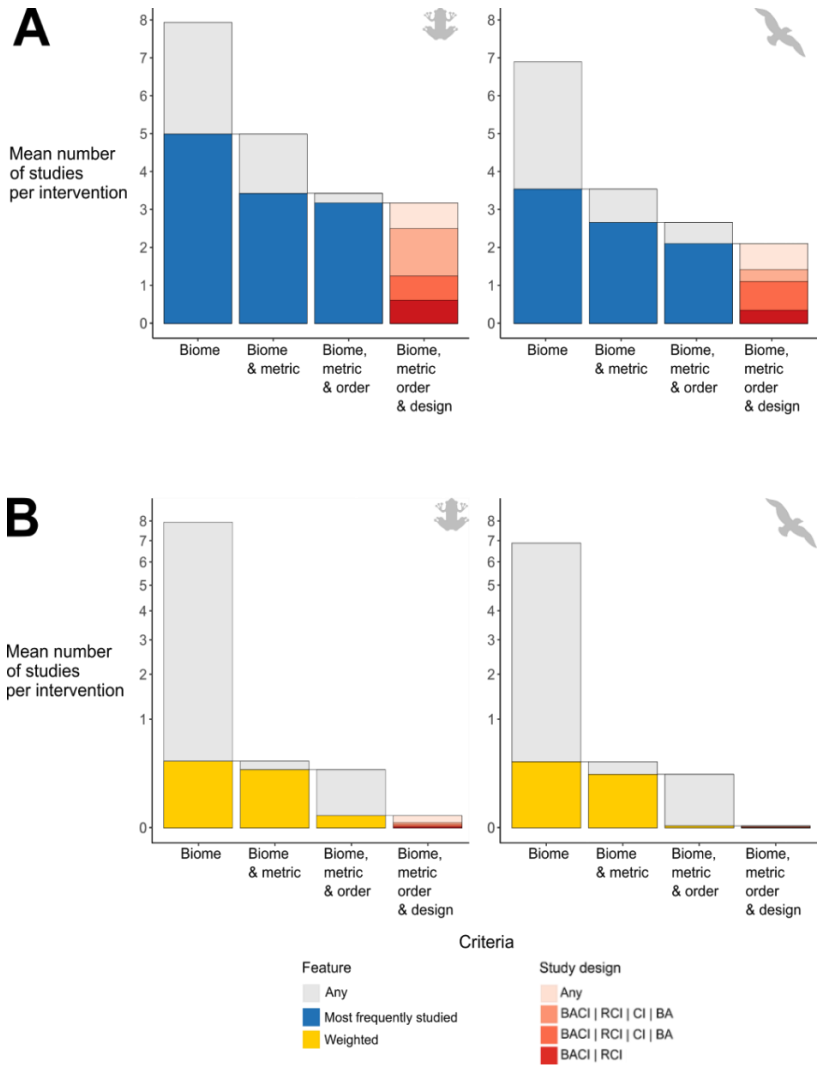


Figure S4 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied biome, metric, and taxonomic order relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies with the most frequently studied biome were carried forward into the biome and metric category. In panel B, studies with a selected biome, metric and taxonomic order were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

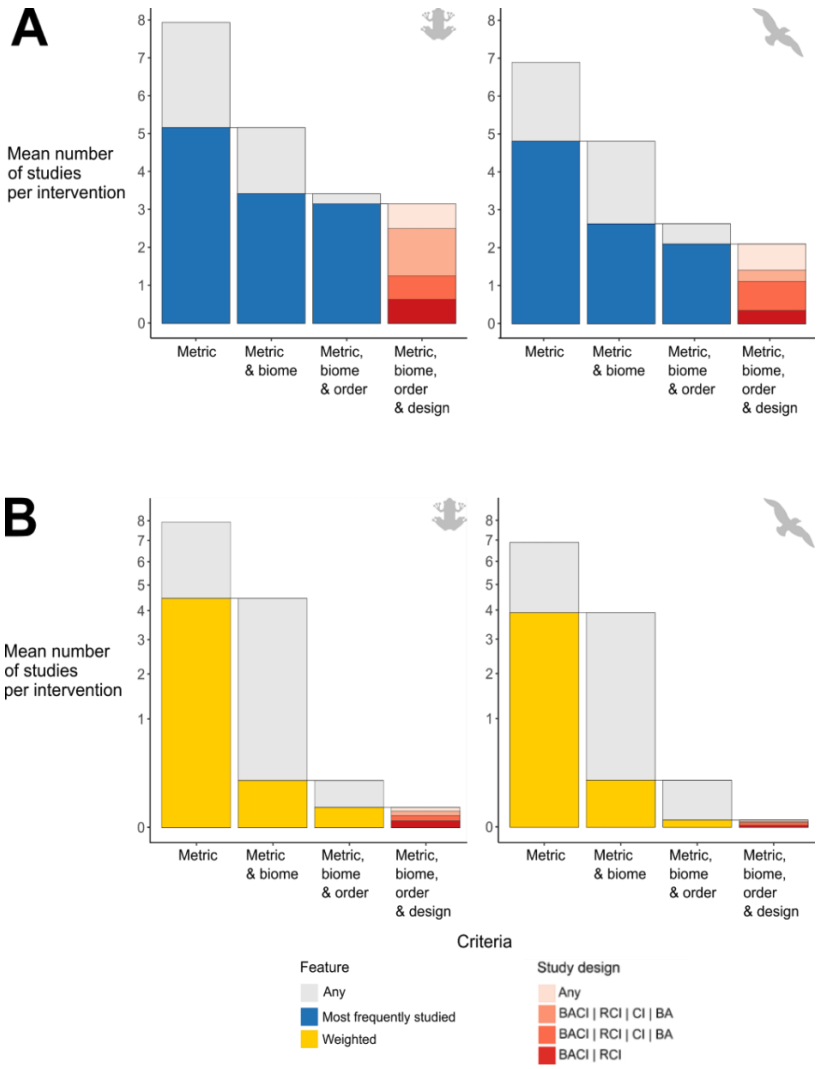


Figure S5 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied metric, biome, and taxonomic order relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies with the most frequently studied metric were carried forward into the metric and biome category. In panel B, studies with a selected metric, biome and taxonomic order were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

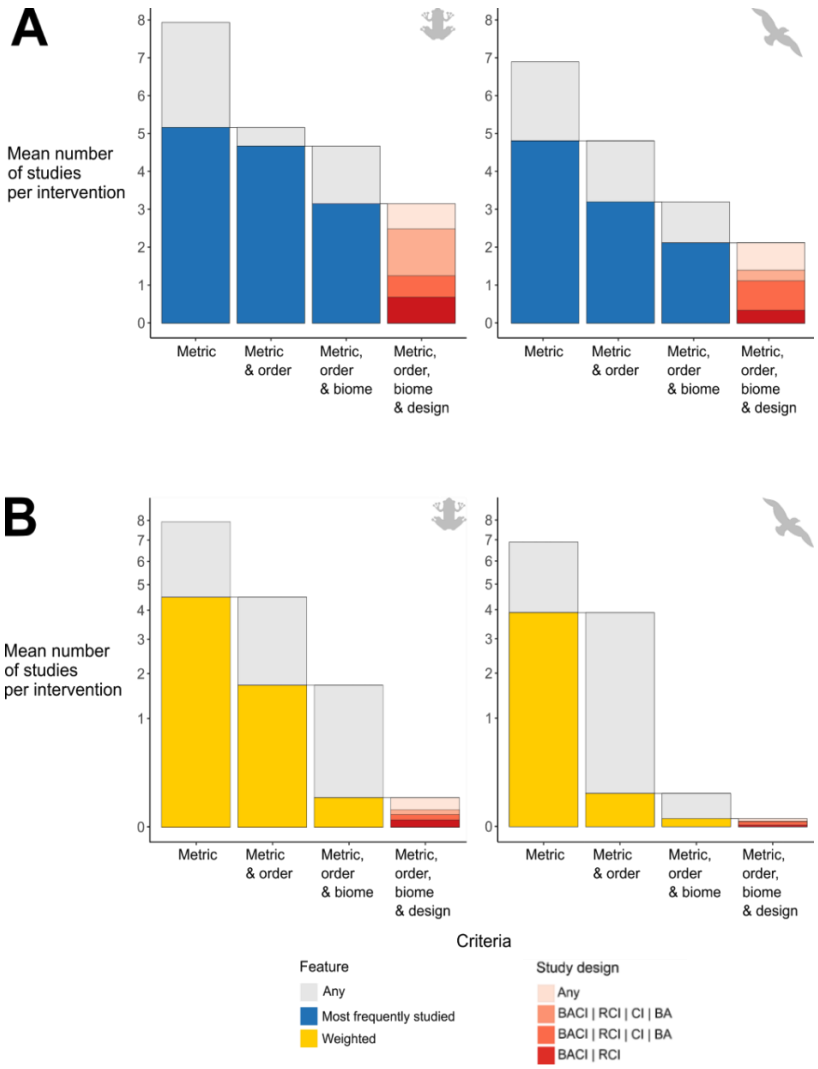


Figure S6 *– Mean numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied metric, taxonomic order, and biome relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies with the most frequently studied metric were carried forward into the metric and order category. In panel B, studies with a selected metric, taxonomic order and biome were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

Figure S7

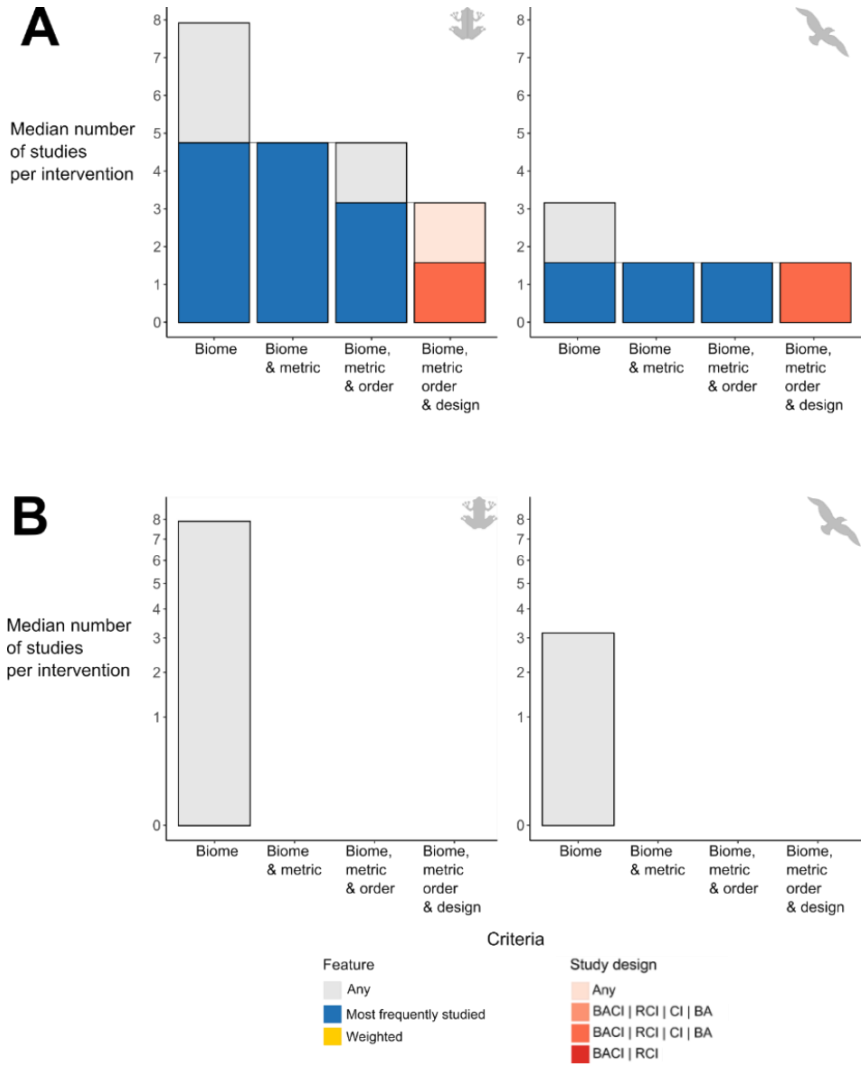


Figure S7 *– Median numbers of amphibian and bird studies per intervention when only considering studies that meet certain relevance criteria. In panel A, studies with the most frequently studied biome, taxonomic order and metric relative to each intervention were counted – here we assume practitioners are interested in the most frequently studied local context. At each step (left to right) we add a further criterion, carrying forward relevant studies from the previous step – for example, only studies conducted in the most frequently studied biome were carried forward into the biome and order category. In panel B, studies with a selected biome, taxon and metric were counted (y axis has a square root transformation). Here we assume practitioners are more likely to be interested in: biomes that are inhabited by higher proportions of threatened species; taxonomic orders that have higher relative proportions of threatened species; and metrics that are most frequently used within each intervention. At the final step, studies are counted based on the study design they use (see Materials and methods for details of study designs).*

Figure S8

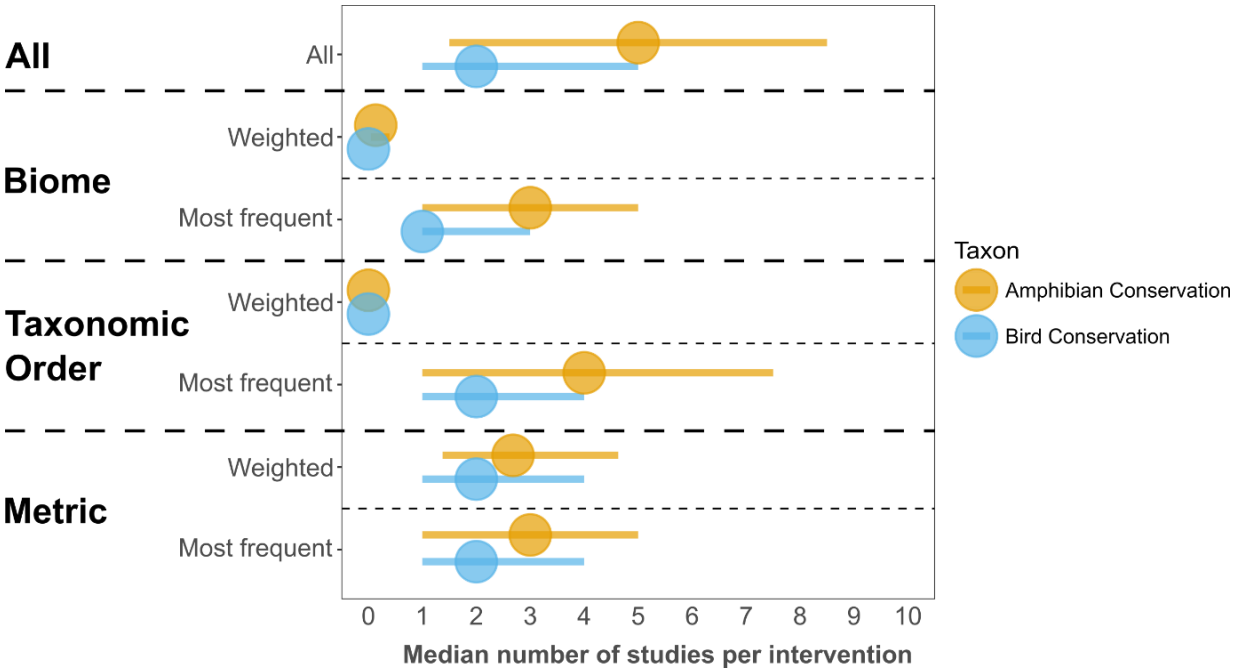


Figure S8 *– Median number of studies per intervention when studies were counted based on whether they considered the most frequently studied biome, metric, or order, and whether they considered a randomly selected biome, metric, or taxonomic order from a weighted list. These weightings were based on the proportion of threatened species found in each biome or taxonomic order. 'All' indicates the median number of studies per intervention when considering all studies.*

Figures S9-S12

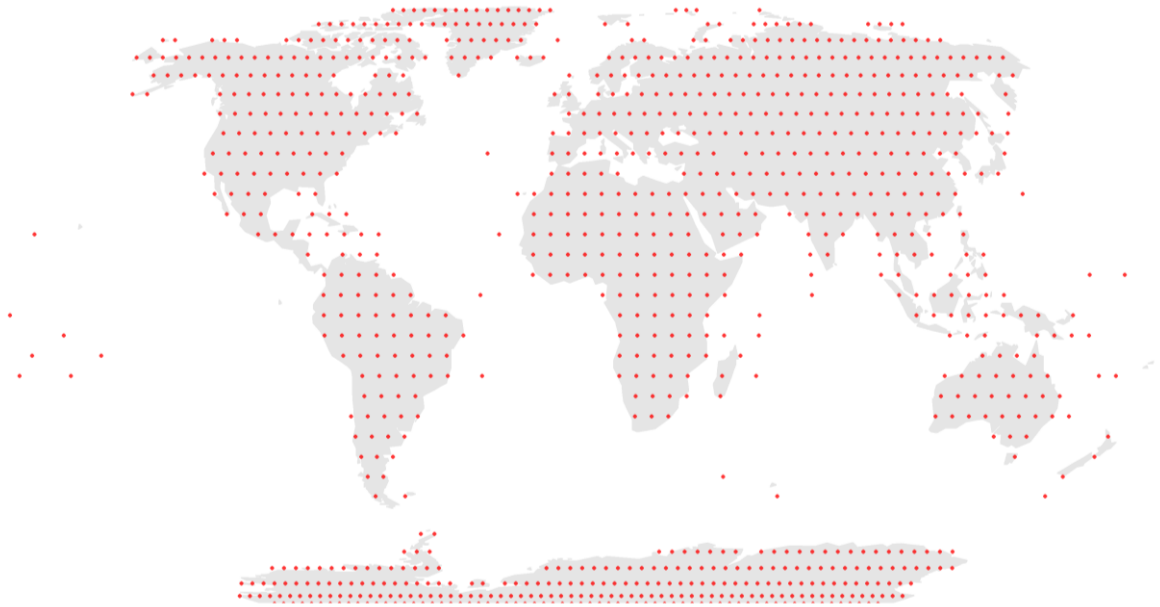


Figure S9 *– Regularly spaced coordinates for birds over terrestrial landmasses with a 1-degree grid cell buffer (OpenStreetMap 2019).*

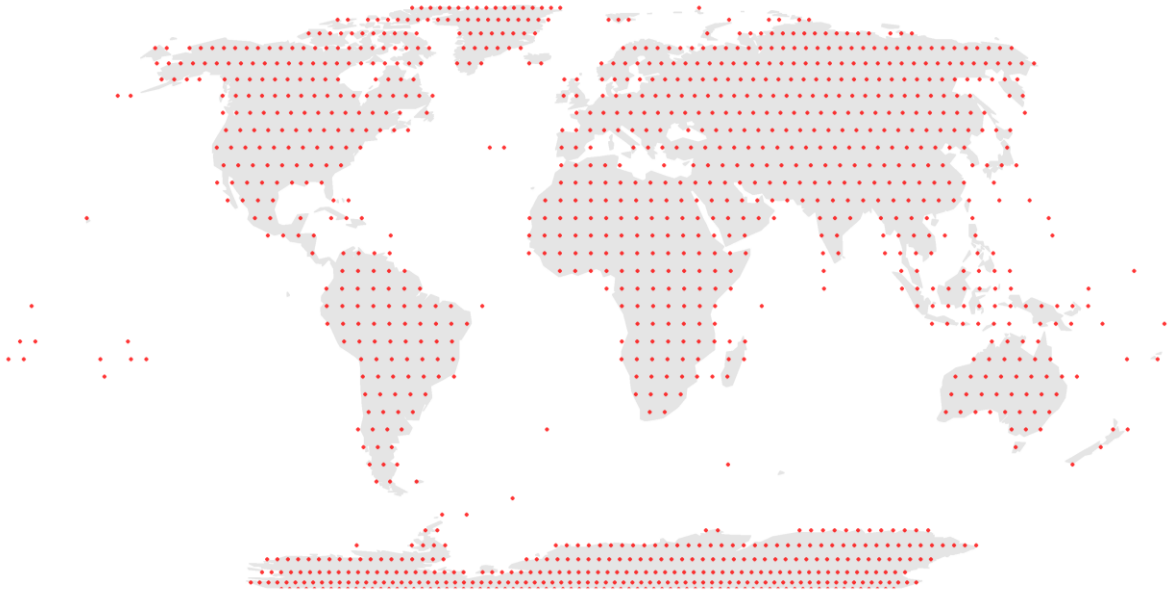


Figure S10 *– Expected coordinates for bird studies if studies were regularly distributed over terrestrial landmasses with a 1-degree grid cell buffer (OpenStreetMap 2019).*

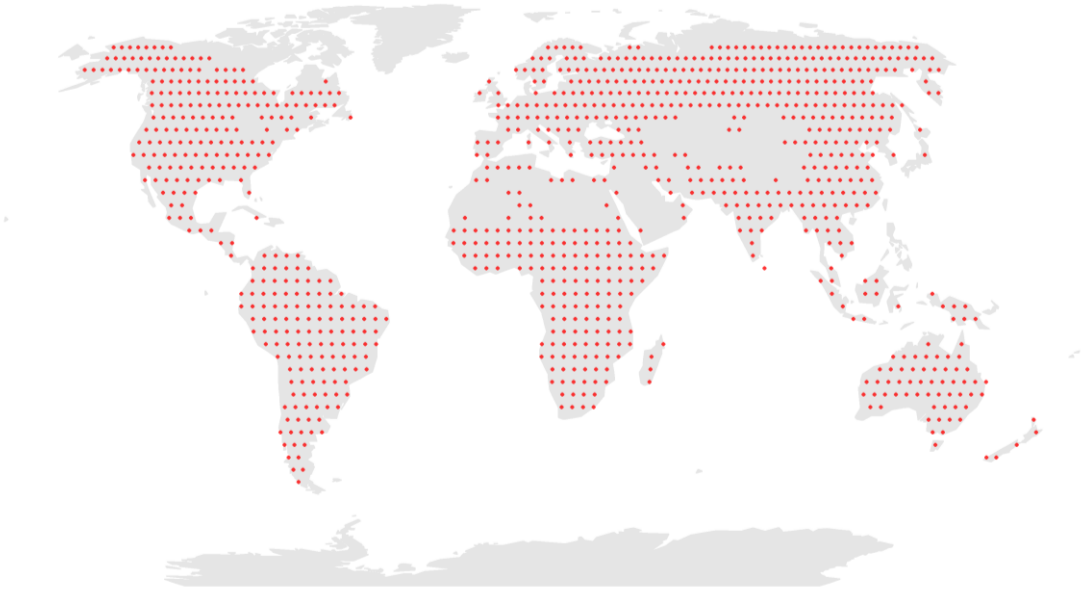


Figure S11 *– Regularly spaced coordinates for amphibians over the combined extent of all amphibian species ranges (IUCN 2019).*

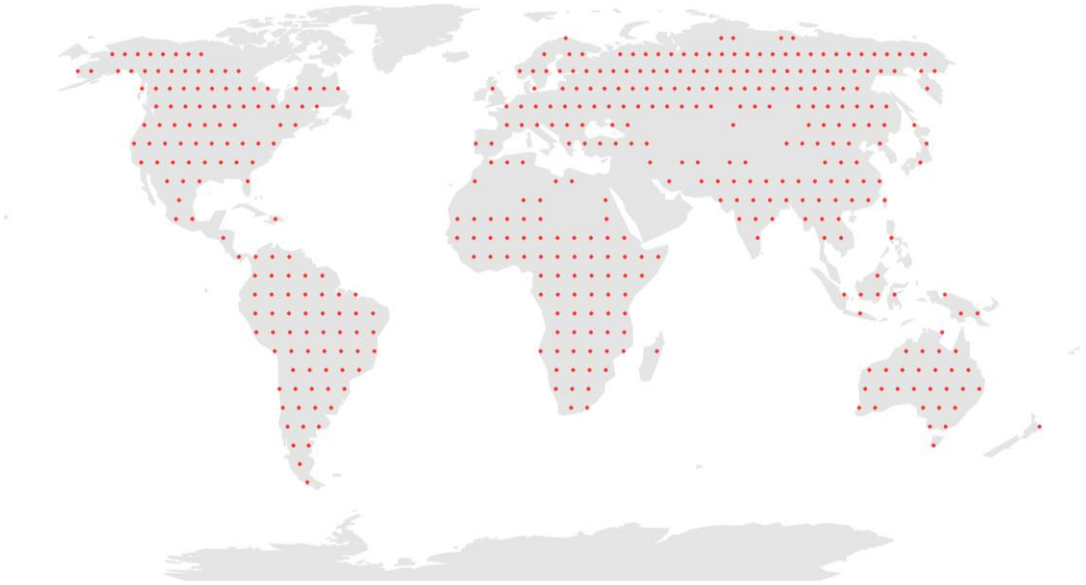


Figure S12 *– Expected coordinates for amphibian studies if studies were regularly distributed over the combined extent of all amphibian species ranges (IUCN 2019).*

Figures S13-S14

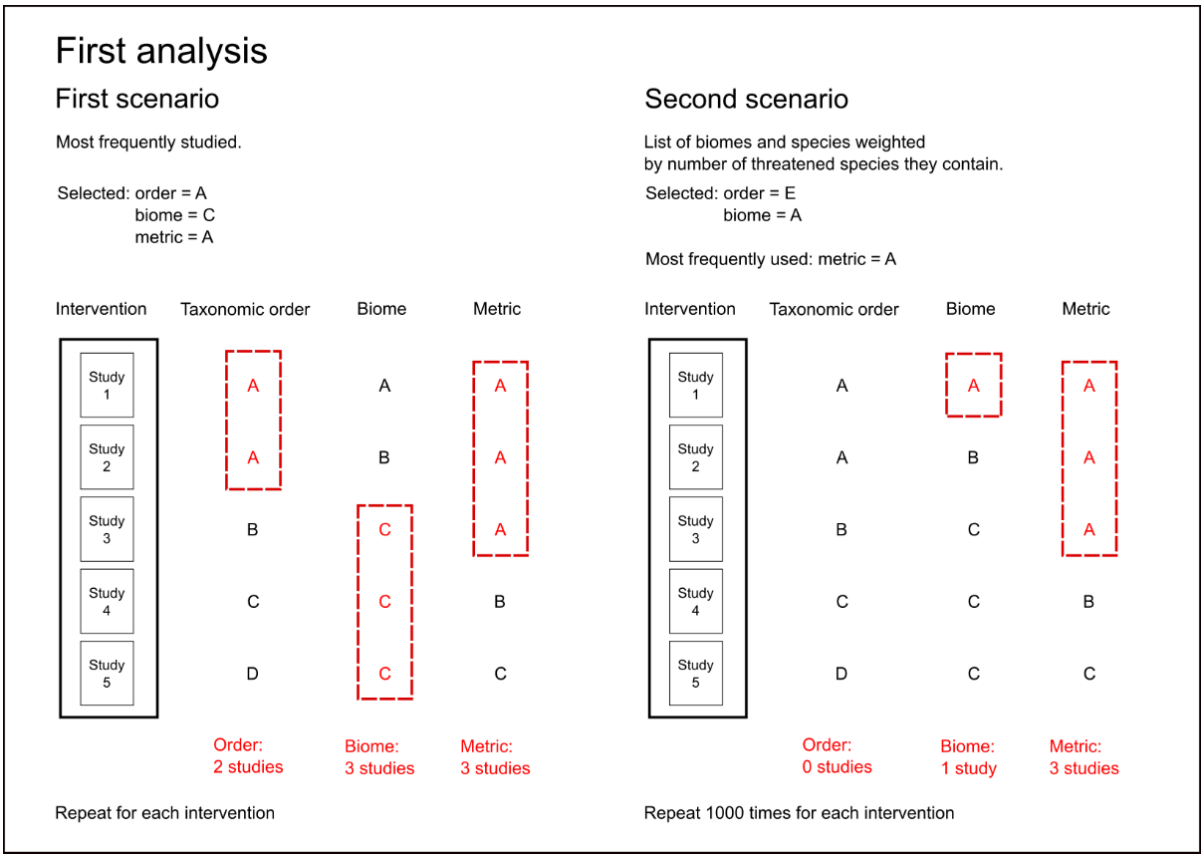


Figure S13 *– Diagram explaining the first analysis described in Materials and methods: Context-specific availability of studies. We show how we measure the number of relevant studies in each intervention using some examples where certain criteria are selected. The first scenario considers the number of studies that are relevant to the most frequently studied local context (i.e., the most commonly studied taxonomic orders, biomes, and metrics). The second scenario considers the number of studies that are relevant to local contexts where the need for conservation is the greatest (i.e., for biomes and taxonomic orders with the most threatened species).*

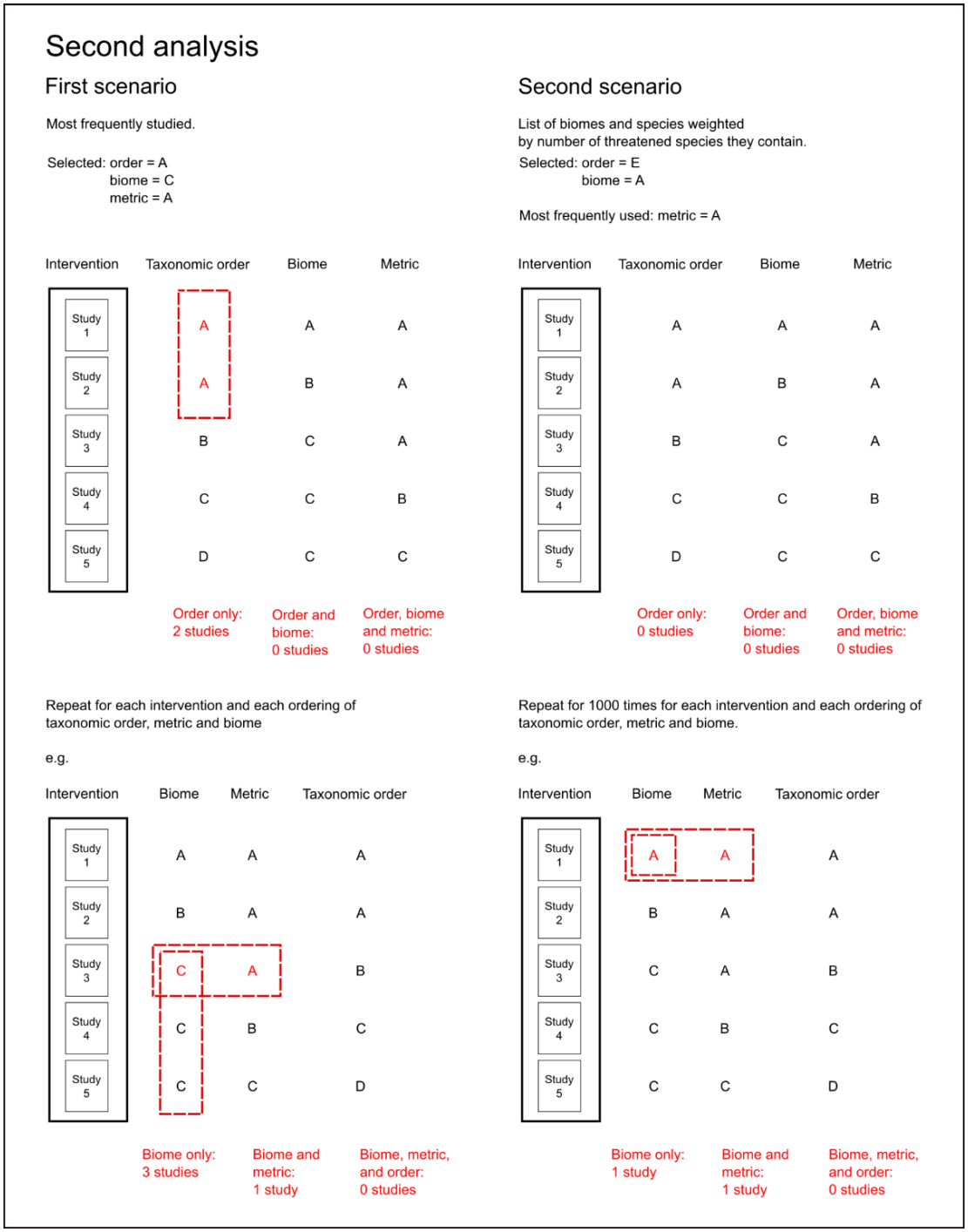


Figure S14 *– Diagram explaining the second analysis described in Materials and methods: Context-specific availability of studies. We show how we measure the number of relevant studies in each intervention using some examples where certain criteria are selected, one at a time in a certain ordering (e.g., first taxonomic order, then biome, then metric). The first scenario considers the number of studies that are relevant to the most frequently studied local context (i.e., the most commonly studied taxonomic orders, biomes, and metrics). The second scenario considers the number of studies that are relevant to local contexts where the need for conservation is the greatest (i.e., for biomes and taxonomic orders with the most threatened species).*

Appendix S1

To obtain the metadata we needed to assess the availability of relevant studies, we used two previously described methods from Christie et al. (2020): i) assigning each study to a biome using coordinates from the Conservation Evidence database, the sp package in R (Bivand et al., 2013; Pebesma and Bivand, 2005), and a shapefile from Dinerstein et al. (2017) (see https://doi.org/10.5281/zenodo.3634780 for R code); and ii) web-scraping of the Conservation Evidence website to obtain the metrics used by each study (see Christie, Amano, Martin, Petrovan, et al. 2020 and https://doi.org/10.5281/zenodo.3634780 for R code to extract metrics). In Christie, Amano, Martin, Petrovan, et al. (2020), we only considered four broad metric types (abundance/density/cover, reproductive success, diversity and survival/mortality), but here we expanded this, extracting 14 different metrics from study summaries on the Conservation Evidence website, which we grouped into the following nine groups: count-based (abundance, density and cover), diversity (diversity and species richness), activity-based (activity, frequency of usage and occupancy), physiological, survival (survival and mortality), reproductive success, education-based, regulation-based, and biomass. Details of keywords used to extract metadata (adapted from Materials and methods in Christie, Amano, Martin, Petrovan, et al. 2020) are found in the R code for extracting metrics available at https://doi.org/10.5281/zenodo.3634780. These reflected similar broad types of metrics that may be used to assess the effectiveness of an intervention to conserve amphibians or birds. The accuracy of metric type extraction from the Conservation Evidence website was 86% from a random 5% of amphibian studies (18 out of 21) and 90% for a random 5% of bird studies (56 out of 62) in the Conservation Evidence database. Accuracy was defined as there being no false positives or false negatives for that study in any intervention that it provided evidence for (as a single study can be found in multiple interventions). Details of false positives and negatives were as follows: for amphibians, three studies with false negatives (failure to detect physiological metric for one study, and failure to detect abundance, density and cover metrics for two studies); for birds, five studies with false positives (incorrectly detected physiological metrics for two studies, reproductive success for one study, survival for one study, and activity-based metrics for one study) and two studies with false negatives (failure to detect activity-based metrics for one study and failure to detect diversity metrics for one study). As this automated classification was used to estimate the mean number of studies per intervention across a large number of interventions for each metric group, these misclassifications will have made little difference to these overall estimates. This automation currently offers the most feasible and reproducible methodology to analyse a large number of studies and controls for some potential biases that would affect manual classification. See main text for references.

6 | Where next for Evidence-Based Conservation?

Introduction

In this final Chapter, I draw together the findings and unifying themes of Chapters 2-5 that investigated two types of biases (within-study and between-study biases) in the evidence base for conservation. This Chapter is structured into sections addressing each theme, in which relevant problems and solutions identified by previous Chapters are discussed, along with their implications for the future of evidence-based conservation (Table 1). Themes are divided into those relating to current issues, which urgently need to be addressed to strengthen the evidence base for conservation, and those relating to future work that is critical to the success of evidence-based conservation. Before these themes are discussed, however, it is useful to briefly summarise the key messages of Chapters 2-5.

Table 1 *– The major themes that tie together Chapters 2, 3, 4, and 5, along with the related key problems and solutions highlighted by each Chapter that will be discussed in due course.*

	Current issues		Future work	
Themes	**Study design matters**	**Gaps and biases in the evidence base**	**Future-proofing evidence synthesis and evidence assessment**	**Tackling the issues of relevance, generalisability and reproducibility**
Chapters	2 3	4 5	2 3	5
Problems	Simpler study designs suffer from quantifably greater levels of bias and yet are commonly used in environmental and social intervention studies.	There are severe geographic, taxonomic, study design-related, and bioclimatic gaps and biases in the evidence base for conservation.	Current evidence synthesis and assessment methods are poorly equipped to cope with the projected growth in studies that need to be synthesised.	Many practitioners will struggle to find relevant evidence due to the mismatch between where actions are tested and where they are needed.
Solutions	Account for study design bias using alternative meta-analytic approaches. Facilitate the use of better study designs.	Prioritised and rigorous testing of actions is needed to address biases. Non-English language and grey literature need to be better integrated.	Embrace Artificial Intelligence and alternative weighting approaches to speed up and automate evidence synthesis and assessment.	Investigate generalisability of conservation actions and develop decision support tools to apply global evidence to local settings.

In Chapters 2 and 3, I focused on investigating within-study biases, whilst in Chapters 4 and 5, I focused on quantifying between-study biases in the literature testing conservation interventions. In Chapter 2, I used empirically driven simulations to show that simpler study designs are typically poor at estimating the true effect of ecological impacts and suggested an alternative method of weighting studies in meta-analysis to account for differences in study design bias. In Chapter 3, I built on Chapter 2 by reanalysing raw experimental and observational data using a more principled, model-based approach to quantify the bias associated with different study designs and how this can be accounted for when synthesising evidence. I also demonstrated that simpler and less credible study designs tend to be more commonly used in conservation science and social science. In Chapter 4, I showed that the evidence base for conservation was highly biased geographically, bioclimatically, and taxonomically, and furthermore that more credible study designs were more geographically

restricted to Western countries. Finally, in Chapter 5, I demonstrated that there is a mismatch between where we test conservation interventions and where they are most needed. Ultimately, this means that many practitioners will struggle to find evidence that they perceive as being relevant to their local setting due to the severe gaps and biases in the evidence base.

Discussion

Study design matters

In Chapters 2 and 3, I highlighted the fundamental importance of within-study biases related to study design in determining the reliability of study findings. I showed that randomised experiments (e.g., the Randomised Control-Impact, RCI, design) and the Before-After Control-Impact (BACI) observational design are clearly worth the extra investment in research effort that they require because they remove many of the biases that affect simpler observational study designs and better account for the stochastic nature of environmental variability (Christie et al., 2020, 2019). In reality, however, simply suggesting that all researchers should only use these less biased study designs is not a helpful recommendation. As I showed in Chapter 3, simpler study designs typically dominate the literature of tests of conservation and social interventions, and I dedicated some discussion to consider why this might be in Chapters 2 and 3. Most notably, the greater use of simpler observational study designs is likely to be due to factors such as: a lack of funding for long-term monitoring; limited statistical education and knowledge on study design; a lack of available data collected before impacts occur; ethical constraints on the use of randomisation; and logistical constraints that mean that researchers do not have advanced knowledge that an impact will occur (Christie et al., 2020, 2019). All these issues can limit the study designs that researchers can realistically use, or the designs they perceive they can use.

To address these barriers, there needs to be concerted action across the scientific research community to enable the use of more rigorous study designs wherever possible. This ranges from the statistical education that researchers at all career stages receive, to the investment of greater resources by institutions and funders to ensure longer-term contracts and funding are available to enable the use of more rigorous designs (Christie et al., 2020, 2019). Pre-registration of studies and pre-analysis plans could also help because the peer review of study designs prior to beginning a study should add an extra stage of quality control to ensure researchers better design their studies (Parker et al., 2019).

In certain situations, we must also accept that no matter what funding or resources are available, it may be impossible to conduct randomised experiments or BACI observational studies. In these situations, I recommended in Chapters 2 and 3 that ecologists and conservationists should embrace the use of other observational study designs that are routinely used in other fields such as economics and epidemiology. Such designs include regression discontinuity designs (Hahn et al., 2001; Maas et al., 2017; Moscoe et al., 2015) to investigate impacts using time series data before and after an impact occurs (a possible alternative to the Before-After design), as well as using instrumental variables (Angrist et al.,

1996) and pairing or matching to strengthen the Control-Impact design (Imbens and Rubin, 2015).

However, I would stress that any cost-based assessment of the feasibility of implementing a particular study design should incorporate the social, environmental, and political costs of Type I and Type II errors associated with different designs (Mapstone, 1995). For example, important interventions or impacts that carry greater risk should warrant the implementation of a higher minimum standard of study design (Mapstone, 1995). Researchers should adjust budgets and project timescales to accommodate study designs, not the other way around.

I believe this thesis has, at the very least, contributed to raising the awareness of the use of different study designs, not only in conservation science, but more widely in the environmental and social sciences. In particular, I have highlighted that it is essential to ensure that studies are designed as rigorously as possible. Building on quotes such as this one by Light et al. (1990): “You can't fix by analysis what you bungled by design...”, I add my own: “Study design is to study, as foundation is to building.” Instilling this mindset in researchers will help to build a stronger and more credible evidence base from which to make more effective decisions in biodiversity conservation and beyond.

Gaps and biases in the evidence base

In Chapters 3 and 4, I demonstrated that the evidence base for conservation is generally of low quality in terms of study design, as well as being extremely patchy and biased taxonomically and geographically. I further highlighted in Chapter 4 that study design and geographic biases are linked because less biased study designs were more restricted to North America, Europe, and Australasia than simpler observational designs. In Chapter 5, I showed that most evidence in conservation is likely to be perceived as being of low relevance to many practitioners, and there is a particular mismatch between where studies are conducted and where the need for conservation action is greatest (i.e., assuming conservationists wish to prioritise conserving threatened species).

These between-study biases are clearly serious and require that urgent action be taken to resolve them. The geographic and taxonomic biases and gaps in the evidence base need to be tackled in a prioritised manner with a joined-up approach across conservation NGOs, practitioners, funders, academics, scientists, and government bodies. It is likely that focusing effort on testing interventions on threatened species will help to resolve both geographic and taxonomic biases simultaneously. Focusing on the tropics would be a good place to start because this region holds a substantial proportion of the world's threatened species (Barlow et al., 2018) and yet is poorly represented in the evidence base. Coordinating such work on testing interventions with the IUCN Species Survival Commission (SSC) groups is likely to be an effective approach to filling these gaps. There is also a case for actively discouraging the testing of certain interventions in well-studied regions, and diverting this research effort to poorly-studied regions; possible ways to implement this could be through funders targeting their investment away from well-studied areas, or through encouraging organisations and researchers to develop new research strategies to prioritise research in poorly-studied areas. However, we should be cautious not to promote more 'Helicopter' or 'Parachute Science', whereby Western researchers have been found to be responsible for a disproportionate amount of research conducted in poorly-studied regions such as Africa, South America, and Asia (Geldmann et al., 2020; Pototsky and Cresswell, 2020). Promoting greater local capacity building and empowering local researchers to conduct and publish tests of interventions will therefore be fundamental to improving the coverage of the evidence base for conservation.

Such work could also remedy some of the geographic bias in the use of study designs; if this is simply due to a lack of studies outside of North America, Europe, and Australasia. It may be, however, that this design-related geographic bias is due to greater logistical constraints in conducting more rigorous study designs outside of these regions where there are fewer resources and less support for research. Greater statistical training and collaborations between researchers, statisticians, and methodologists from different countries will help to

build greater capacity in underrepresented regions to not only conduct tests of interventions, but also ensure that they are designed robustly.

One limitation of the Conservation Evidence database that formed the basis for much of my analyses in this thesis is that it is, at present, mostly composed of English language studies. This limitation, however, provides hope that the between-study biases in the evidence base for conservation can be mitigated through incorporating more tests of interventions from the non-English language literature. Preliminary work I have undertaken in collaboration with Dr Tatsuya Amano and the Transcending Language Barriers to Environmental Science (translatE) project suggests that non-English language studies could increase the taxonomic coverage of Conservation Evidence by 47% (Amano and Espinola, 2020a) and increase the geographic coverage by 7% (Amano and Espinola, 2020b). However, this work also suggests that simpler, less credible study designs are also used more often in non-English language studies. This further highlights the need for greater statistical training and collaborations in a range of locations and languages to improve the use of more credible study designs outside of North America, Europe, and Australasia where almost all non-English language studies are conducted.

Of course, these efforts to integrate more non-English literature into the predominantly English language evidence base only address one side of the language bias coin. We also need to consider how to reverse this transfer of knowledge from English to non-English languages. Many conservation practitioners around the world do not speak English and so may have limited access to knowledge on what does and does not work in conservation that is predominantly only available in English (Amano et al. 2016). Therefore, efforts to provide non-English language translations of study summaries in evidence databases such as Conservation Evidence (as well as study abstracts in journals) is an important endeavour. Embracing improvements in translations provided by Artificial Intelligence (AI) (King, 2019) and forming more global, international collaborations between researchers in English-speaking and non-English speaking countries will help to ensure we can make evidence-based conservation a globally successful movement.

Grey literature could also hold some promise to filling in some of the current knowledge gaps in the evidence base for conservation. This grey literature, typically defined as studies that have not been published in a peer-reviewed journal, is often composed of reports by organisations and government bodies, theses, dissertations, and newsletters (Haddaway and Bayliss, 2015). The rigour of these studies is often highly variable and likely to be less rigorous than the peer-reviewed literature (Haddaway and Bayliss, 2015). Integrating these studies is likely to improve the geographic and taxonomic coverage of the evidence base, but may also

add many poorly designed studies. Integrating grey literature and non-English literature into the evidence base must therefore be carefully conducted to ensure that the rigour of evidence syntheses is upheld – for example, ensuring that rigorous critical appraisal of studies is conducted (Haddaway and Bayliss, 2015). However, I would argue that the risk of ignoring evidence from the grey and non-English literature seems far greater to the wider goals of evidence-based conservation efforts, particularly with careful and considered integration of this literature, than the potential risk of integrating studies with lower study quality.

Deciding on an optimal strategy to fill knowledge gaps and biases ultimately links to the trade-off between internal and external validity. Internal validity is strongly linked to within-study biases because internal validity is the overall quality of a study within that study's setting, whilst external validity reflects how well study findings generalise to other settings (Mupepele et al., 2016). The trade-off between these two forms of validity in the setting of evidence-based conservation can be described as a resource allocation issue. For example, with limited research effort, what is the best strategy to build and strengthen the evidence base for conservation?

We can strengthen the evidence base in two ways: increase its coverage (and so its relevance and external validity) or increase its quality (in terms of the internal validity of studies). To increase coverage, the optimal strategy would be to prioritise tests of interventions using less credible study designs that are cheaper and easier to implement, allowing us to test more interventions across more local settings. To increase quality, the optimal strategy would be to focus on rigorously evaluating fewer interventions across fewer settings using more credibly designed studies.

From the findings of Chapters 2 and 3, there is a danger that focusing on increasing coverage could lead to misinforming decision-makers with evidence drawn from studies that suffer from study design biases. Rather than increasing the efficiency and effectiveness of conservation practice, this could inadvertently have the opposite effect and erode trust in evidence-based conservation. As Wauchope (2020, p127-128) argued, acting based on some misleading evidence could be worse than not acting at all. Iacona et al. (2017) also argued that delaying action for the right amount of time, such as by waiting for stronger evidence, can be an optimal strategy in a crisis discipline and enable conservationists to protect more species more quickly.

Another element to consider in this debate is the need to test for consistency in study findings in terms of reproducibility (testing within the same setting; Begley and Ioannidis, 2015) and generalisability (testing across different settings; Kneale et al., 2019). To test reproducibility, we need to conduct studies in the same or similar study setting using the same experimental or observational design. Study design is linked to reproducibility because if study design bias

is large and inconsistent, different results may be obtained by repeating the same study (Munafò et al., 2017). To test reproducibility, it is therefore sensible to repeatedly test an intervention using a few studies in the same setting with an identical and highly credible study design. To test generalisability, we need to maximise the number of studies testing an intervention in different settings, which would suggest we should focus on conducting many, less credibly designed studies. However, if studies use different designs which tend to be less credible, it may be difficult to disentangle differences in study results from study design bias and genuine differences in the effectiveness of an intervention in different study settings.

Increasing the coverage of the evidence base without considering the individual quality of studies being conducted also risks artificially inflating the confidence of decision-makers in the conclusions drawn from that evidence base. For example, vote-counting (Hedges and Olkin, 1980) is a known problem (whereby tallies are made of the number of studies providing positive or negative results to come to a decision) that does not take account of the internal validity of studies. Such issues with the interpretation of the evidence base makes it even more important to ensure that the findings of less credibly designed studies are appropriately caveated with greater uncertainty.

Above the level of prioritising the type of study designs that are implemented to test an intervention, there is also a dilemma over which types of interventions should be tested in the first place. To maximise the aggregation of marginal gains, a concept that could deliver considerable benefits to conservation (Sutherland, 2019), the testing of interventions with a small number of studies should clearly be prioritised over those with many studies. However, what should we do when the choice is between prioritising interventions with little evidence versus interventions with no evidence? Arguably, the addition of a single study to interventions with no evidence at all is not likely to improve our certainty in the overall evidence base by a large amount – there will still be large uncertainty over the effectiveness of these interventions. Alternatively, the addition of a single study to interventions with some limited evidence would likely have a greater overall benefit for the evidence base because our certainty in the effectiveness of these interventions will be raised to a much more acceptable level. Once an acceptable level of certainty has been reached for these interventions, testing interventions with no evidence would then become a priority.

Ultimately, the most sensible strategy to prioritising research effort to strengthen the internal and external validity of the evidence base for conservation will be a mixed one linked to the prioritisation of the intervention being tested. Prioritising interventions based on additional factors other than simply the current number of studies available is likely to be most effective to achieve the broader goals of evidence-based conservation. For example, priority

interventions could be identified as those that are likely to be highly costly or risky, or those that are commonly used, which could therefore represent a large waste of resources if that intervention is shown to be ineffective. It also seems sensible to prioritise interventions that are designed to conserve threatened species and habitats, given that these could have the greatest potential impact to bend the curve of biodiversity loss (Leclère et al., 2020).

Once high priority interventions are identified, a sensible strategy would be to test their effectiveness using credibly designed studies to limit the chance that costly, misleading conclusions are drawn from the evidence base. For less important, less risky interventions, it may be sensible to prioritise expanding the coverage of the evidence base in the short-term, prioritising credibly designed studies where possible whilst acknowledging that this may not always be possible, and that less credibly designed studies carry additional levels of uncertainty. Ultimately, the choice of research design will lay in the hands of the researcher but could be strongly influenced by funders, journals, and the scientific community, particularly for high priority interventions. A structured approach would be useful to identify and prioritise the evidence that needs to be accumulated to better inform decision-makers for different interventions. For example, it will be important to specifically identify and disseminate knowledge gaps for different conservation interventions, building on the work I presented in Chapter 4, and then to conduct studies to fill these gaps based on the level of priority assigned to each intervention.

At present, this kind of prioritisation approach does not seem to be widely used in conservation, if at all. However, such an approach will be key to ensuring that funders, organisations, and research institutions facilitate the growth and strengthening of the evidence base for conservation in an efficient, timely, and useful manner. Therefore, creating a protocol and standardised approach for prioritising the future testing of conservation interventions, potentially by learning from the IUCN Red List approach to categorising species by risk of extinction (Vié et al., 2009), seems to be urgently needed to ensure evidence-based conservation itself is as efficient as possible.

Future-proofing evidence synthesis and evidence assessment

I believe that evidence synthesis is poorly equipped to cope with the projected rapid growth in the publication of scientific evidence. Evidence synthesis needs to become better prepared for the future, and this needs to occur urgently and efficiently, to keep pace. Conventional systematic reviews and meta-analyses are time-consuming (Haddaway and Westgate, 2019) and quickly go out of date as new studies are published – therefore, new approaches to evidence synthesis need to be embraced (Shackelford et al., 2021; Thomas et al., 2017). AI and Machine Learning hold considerable promise to dramatically speed up the process of identifying relevant literature for evidence syntheses, which often constitutes a substantial proportion of the time taken to conduct systematic reviews and subject-wide evidence synthesis (Cornford et al., 2021; Wallace et al., 2014).

Living systematic reviews have also been proposed, which would exist online and be dynamically updated with new studies as they are published (Thomas et al., 2017; Tsafnat et al., 2013). The Metadataset project (Metadataset, 2020) also provides a basis from which to design dynamically updated and interactive systematic reviews and meta-analyses (Shackelford et al., 2021). The idea of dynamic meta-analyses is something that would be best supported by subject-wide evidence synthesis (Sutherland et al., 2019). As I discussed in Chapter 1, subject-wide evidence synthesis ensures that a dynamically updated database of studies can be collated, albeit at great cost in the short term, to provide significant long-term benefits from economies of scale (Sutherland et al., 2019). This process also enables several different research, policy, and practice questions to be answered as and when required. This is because once the living database has been established, studies, summaries, and assessments can be easily be retrieved rather than having to be repeatedly searched for, extracted, assessed, and summarised for each research question.

Furthermore, advances in AI and Machine Learning could effectively lead to an entirely automated, or at least semi-automated, system of evidence synthesis by computerising both the extraction of relevant studies and their results (Marshall et al., 2020; Marshall and Wallace, 2019; Tsafnat et al., 2013). Private ventures, such as Semantic Scholar (Fricke, 2018) have pioneered the creation of a tool that summarises scientific studies (SCITLDR) into summaries called TLDRs (inspired by the social media acronym that stands for: Too Long; Didn't Read). Such approaches are in their infancy but provide one step towards automating the entire process of evidence synthesis.

Of course, key processes that may be more challenging to automate are evidence assessment and critical appraisal, which often comprise a substantial, but extremely important phase of evidence synthesis (Marshall et al., 2015). The difficulty in doing this is that greater speed

could come at the cost of reduced rigour and the risk of misinforming decision-makers. Therefore, the major challenge is for the evidence synthesis community to come together to understand how automation can be used to speed up evidence synthesis, evidence assessment, and critical appraisal without compromising their integrity – and of course its perceived integrity by decision-makers. The idea of an alternative weighting approach that I outlined in Chapter 2, and made more robust, model-based, and principled in Chapter 3, provides a basis from which to design more advanced, automated systems of weighting and critical appraisal that incorporate a wider range of valid indicators of study quality. Such approaches will require more formalised, structured ways of presenting the methodology, study design, data, and results, such as those used in medicine to report the results of drugs trials (Moher et al., 2001). Detailing key information in ways that enable easy, automated extraction will be key; for example, automatic weighting approaches for evidence assessment and meta-analysis would need to automatically extract the results of studies and the key indicators of study quality that determine study weights. In-depth testing, innovative methods drawing upon advances in AI and Machine Learning, as well as clear communication and transparency behind these automated processes will be key to successfully future-proofing evidence synthesis.

Finally, it would also be beneficial to ensure that tests of conservation interventions become more standardised to aid the speed and automation of future evidence syntheses. When different metrics are used, it becomes difficult to conduct syntheses, particularly quantitative meta-analyses, as studies may be measuring the effectiveness of interventions on fundamentally different outcomes (e.g., the diversity of birds or abundance of one species). As Chapter 5 demonstrated, the inconsistency in the use of metrics often means that there are few studies that will be directly relevant to a decision-maker if they desire evidence on a certain aspect of the effectiveness of an intervention. Of course, using multiple metrics is valuable to understanding different aspects of an intervention's effectiveness, but to aid evidence synthesis these need to be used consistently by studies testing the same intervention. Therefore, agreeing on and disseminating a list of suitable metrics with which to test different interventions would help to ensure evidence is being generated in an efficient and organised manner. Such work should be prioritised for important interventions, as discussed earlier, and could be coordinated with practitioners and IUCN SSC groups to ensure researchers understand the need to use recognised and standardised measures of interventions' effectiveness.

Tackling the issues of relevance, generalisability, and reproducibility

Considering the future of evidence-based conservation would not be complete without further exploring issues surrounding the generalisability and reproducibility of research findings. Ultimately, the aim of evidence-based conservation is to inform and improve conservation practice and policy by enabling informed decision-making using evidence. This evidence is often summarised at a global and highly generic level in evidence summaries, systematic reviews, and meta-analyses (Nakagawa et al., 2020). Indeed, finding generality in patterns, laws, and processes is a particular motivation for scientific researchers (Lawton, 1999). However, decision-makers often think and operate at local scales and typically prefer to only consider evidence that they perceive as being locally relevant to them (Gutzat and Dormann, 2020). As a result, evidence syntheses that summarise evidence at a global, generic level have been criticised for lacking realism and providing a 'view from nowhere' (Shapin, 1998). These issues, coupled with the patchiness of the evidence base for conservation that I highlighted in Chapters 4 and 5, pose a fundamental problem for evidence-based conservation and evidence synthesis. Namely, how do we apply the global, patchy evidence base to different decision-makers' local settings?

This issue is directly related to the generalisability of research findings because to go from global evidence to local questions, we need to understand how applicable and transferable research findings are between different settings (Kneale et al., 2019). This issue was first formally discussed by Levins in the context of mathematical models for complex biological systems, where he defined a trade-off between realism, precision, and generality (Levins, 1968, 1966). Scientists working in evidence synthesis generally focus on precision and generality when collating evidence at a global scale, whilst local decision-makers may often focus on realism. Reconciling these differences in approaches is a fundamental problem for all of science, but particularly for evidence-based conservation where we know that context-dependency of research findings due to socioecological, bioclimatic, and taxonomic factors can be substantial (Finch et al., 2019; Shapin, 1998). Better understanding the extent to which research findings generalise across local contexts is therefore crucial to ensuring evidence syntheses can provide locally relevant recommendations to decision-makers. In the future, I envisage that we could produce evidence-based guidelines that categorise conservation interventions into several broad categories (Fig.2) based on two variables: generalisability and effectiveness. For example, can we identify and promote conservation interventions that are universally effective? And can we advocate against using interventions that are universally ineffective? For interventions that appear to be highly context-dependent, can we prioritise research to investigate and predict under which circumstances these interventions are effective?

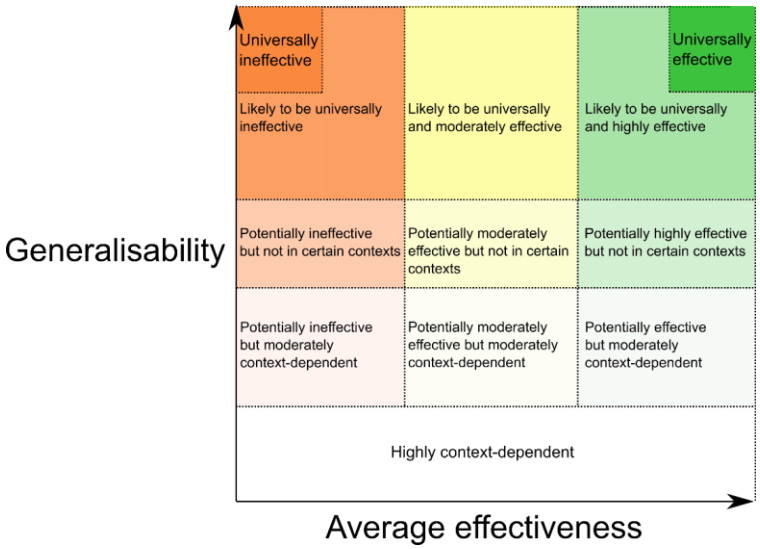


Figure 2 *– A possible matrix for categorising interventions by their average effectiveness across study settings and generalisability to produce evidence-based guidance.*

Reproducibility is another issue that warrants more attention in evidence-based conservation. Its recent rise to prominence in several disciplines occurred after several large analyses drew attention to the lack of replicability in study findings in economics, clinical science, and psychology (Begley and Ioannidis, 2015; Errington et al., 2014; Munafò et al., 2017; Open Science Collaboration, 2015). Reproducibility is also intrinsically linked to study design since using less biased study designs should help researchers to produce more reliable and repeatable results (Munafò et al., 2017). Many other factors, some of which are within-study biases that I was unable to investigate in this thesis, affect reproducibility and merit greater investigation in conservation. For example, Questionable Research Practices (QRPs) is a term used to describe factors including HARKing (Hypothesising after Results are Known) and p-hacking, and it is important that their prevalence and severity in conservation is quantified (Munafò et al., 2017). A project I am involved with led by Dr Hannah Fraser (the 'Same Data Different Analysts' project; Fraser et al., 2020) is tackling part of this issue by examining how different modelling approaches may lead to different study findings and conclusions. Comprehensively quantifying and improving reproducibility in conservation will also help to strengthen the rigour of (and decision-makers' trust in) the scientific evidence base.

Addressing relevance, generalisability, and reproducibility in conservation will undoubtedly be a major challenge given the biased nature of the evidence base for conservation. However, I believe there are several tools that could be effectively used in future to investigate these issues. First, to quantify reproducibility, analyses could be undertaken of conservation interventions tested by many studies that act as 'direct' or 'partial replications' to investigate the extent to which study findings replicate under similar study conditions. In the same analysis, generalisability could also be quantified by comparing between-study differences in

findings and modelling factors that may predict these differences (e.g., quantitative traits, habitats, climatic variables, sociological factors). Decision-makers could also be used in these analyses by using structured elicitation approaches, such as a modified Delphi technique using the IDEA protocol (Hemming et al., 2018), to elicit expert judgements on how likely study findings are to replicate or generalise across different local settings and contexts. Furthermore, Prediction Markets (DellaVigna et al., 2019) could be used to estimate the generalisability of study findings where 'forecasters' are asked to predict study effect sizes for different hypothetical study settings. This has been successfully used in the social sciences, where tests have shown forecasters can achieve remarkably high levels of accuracy (DellaVigna et al., 2019). If generalisability and relevance of research can be quantified objectively, it could add rigour to approaches such as dynamic meta-analysis (implemented as part of the Metadataset (2020) project on www.metadataset.com), which seeks to weight study results by their relevance to the decision-maker (Shackelford et al., 2021), as well as measures of study quality (such as using the approaches I have proposed in this thesis). Weighting by relevance in meta-analysis, and more widely in evidence assessment, would make evidence synthesis more customisable and directly apply the global evidence base to the local setting of interest to a given decision-maker ('bringing meta-analysis to the masses').

Ultimately, we also need to better understand how practitioners use evidence in conservation and how we can best nudge and guide decision-makers to adopt evidence-based approaches to their decision-making. An issue that has struck me during my PhD is how little evidence we have on how decision-makers actually use evidence and how to combine diverse sources of knowledge and evidence (e.g., scientific evidence and local knowledge). This is a fundamental barrier at the heart of research-practice-policy gaps in conservation. I have co-designed a decision-support tool during my PhD studentship that guides practitioners through the process of making an evidence-based decision (the Evidence-to-Decision tool: www.evidence2decisiontool.com) based on the Evidence-to-Decision framework (Alonso-Coello et al., 2016) used in healthcare by the National Institute for Health and Care Excellence (NICE). It will be extremely important for more researchers in the evidence synthesis and conservation science community to proactively engage and involve practitioners in the co-design of decision-support tools to embed evidence-based decision-making into practitioners' organisational structures and processes.

Conclusion

Evidence-based conservation has come a long way since the seminal paper by Sutherland et al. (2004) that catalysed the development of the Conservation Evidence project and Collaboration for Environmental Evidence, amongst other conservation-related evidence synthesis projects. Much progress has been made in applying many of the approaches used in evidence-based medicine, such as systematic reviews, meta-analyses, and critical appraisal, as well as some new approaches, such as subject-wide evidence synthesis, dynamic meta-analysis, and weighting by study quality and relevance. My work has shone a light on how, despite accumulating a large database of studies testing conservation interventions, our knowledge of what does and does not work in conservation is still patchy and suffers from severe biases.

My thesis has identified key knowledge gaps and biases that we can learn from to strengthen and improve the evidence base for conservation. This ranges from improving the use of more rigorous study designs, to testing more conservation interventions in underrepresented locations and on underrepresented species. My thesis has also drawn attention to wider study design-related biases that not only affect evidence-based conservation, but also the reliability of study findings across science more widely. I have shown that we can not only quantify the relative magnitude of study design bias affecting studies but also go further by explicitly accounting for design bias using alternative weighting approaches in evidence synthesis. My thesis therefore provides a springboard towards ensuring evidence-based conservation meaningfully delivers more effective conservation practice.

Based on the results of my thesis, my first recommendation is that evidence-based conservation urgently strengthens and expands its evidence base through greater collaboration between researchers, methodologists, and practitioners around the world to fill the knowledge gaps and biases I have quantified. Second, researchers in evidence-based conservation need to engage more strongly with decision-makers to understand their needs and how to make evidence synthesis products more relevant to them, such as through the co-design of decision-support tools. And third, researchers should embrace new technologies and automation that could help to future-proof evidence synthesis to meet the challenges of collating, assessing, and disseminating a rapidly growing evidence base. The future success of biodiversity conservation depends on evidence-based conservation; if evidence-based conservation fails, conservation is unlikely to become efficient or effective enough to tackle the biodiversity crisis. I hope my thesis will contribute to the future success of biodiversity conservation and inspire others to work towards this critically important goal.

References

Alonso-Coello, P., Oxman, A.D., Moberg, J., Brignardello-Petersen, R., Akl, E.A., Davoli, M., Treweek, S., Mustafa, R.A., Vandvik, P.O., Meerpohl, J., Guyatt, G.H., Schünemann, H.J., 2016. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. British Medical Journal 353, i2089. https://doi.org/10.1136/bmj.i2089

Amano, T., Espinola, V.B., 2020a. December 2020 update on the progress of translatE project. https://translatesciences.com/wp-content/uploads/2020/12/Update181220.pdf

Amano, T., Espinola, V.B., 2020b. October 2020 update on the progress of translatE project. https://translatesciences.com/wp-content/uploads/2020/10/Update191020.pdf

Amano, T., González-Varo, J.P., Sutherland, W.J., 2016. Languages Are Still a Major Barrier to Global Science. PLOS Biology 14, e2000933. https://doi.org/10.1371/journal.pbio.2000933

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–455. https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902

Barlow, J., França, F., Gardner, T.A., Hicks, C.C., Lennox, G.D., Berenguer, E., Castello, L., Economo, E.P., Ferreira, J., Guénard, B., Gontijo Leal, C., Isaac, V., Lees, A.C., Parr, C.L., Wilson, S.K., Young, P.J., Graham, N.A.J., 2018. The future of hyperdiverse tropical ecosystems. Nature 559, 517–526. https://doi.org/10.1038/s41586-018-0301-1

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in Science. Circulation Research 116, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Christie, A.P., Abecasis, D., Adjeroud, M., Alonso, J.C., Amano, T., Anton, A., Baldigo, B.P., Barrientos, R., Bicknell, J.E., Buhl, D.A., Cebrian, J., Ceia, R.S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M.D., Davoult, D., de Backer, A., Donovan, M.K., Eddy, T.D., França, F.M., Gardner, J.P.A., Harris, B.P., Huusko, A., Jones, I.L., Kelaher, B.P., Kotiaho, J.S., López-Baucells, A., Major, H.L., Mäki-Petäys, A., Martín, B., Martín, C.A., Martin, P.A., Mateos-Molina, D., McConnaughey, R.A., Meroni, M., Meyer, C.F.J., Mills, K., Montefalcone, M., Noreika, N., Palacín, C., Pande, A., Pitcher, C.R., Ponce, C., Rinella, M., Rocha, R., Ruiz-Delgado, M.C., Schmitter-Soto, J.J., Shaffer, J.A., Sharma, S., Sher, A.A., Stagnol, D., Stanley, T.R., Stokesbury, K.D.E., Torres, A., Tully, O., Vehanen, T., Watts, C., Zhao, Q., Sutherland, W.J., 2020. Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. Nature Communications 11, 6377. https://doi.org/10.1038/s41467-020-20142-y

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I., Sutherland, W.J., 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56, 2742–2754. https://doi.org/10.1111/1365-2664.13499

Cornford, R., Deinet, S., de Palma, A., Hill, S.L.L., McRae, L., Pettit, B., Marconi, V., Purvis, A., Freeman, R., 2021. Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. Global Ecology and Biogeography 30, 339–347. https://doi.org/https://doi.org/10.1111/geb.13219

DellaVigna, S., Pope, D., Vivalt, E., 2019. Predict science to improve science. Science 366, 428–429. https://doi.org/10.1126/science.aaz1704

Errington, T.M., Iorns, E., Gunn, W., Tan, F.E., Lomax, J., Nosek, B.A., 2014. Science forum: An open investigation of the reproducibility of cancer biology research. eLife 3, e04333. https://doi.org/10.7554/eLife.04333.001

Finch, T., Branston, C., Clewlow, H., Dunning, J., Franco, A.M.A., Račinskis, E., Schwartz, T., Butler, S.J., 2019. Context-dependent conservation of the cavity-nesting European Roller. Ibis 161, 573–589. https://doi.org/10.1111/ibi.12650

Fraser, H., Parker, T.H., Nakagawa, S., Fidler, F., Gould, E., Gould, E., Vesk, P.A., Hamilton, D.G., 2020. Many EcoEvo Analysts [WWW Document]. OSF. URL osf.io/mn5aj

Fricke, S., 2018. Semantic Scholar. JMLA 106, 145–147. https://doi.org/10.5195/jmla.2018.280

Geldmann, J., Alves-Pinto, H., Amano, T., Bartlett, H., Christie, A.P., Collas, L., Cooke, S.C., Correa, R., Cripps, I., Doherty, A., Finch, T., Garnett, E.E., Hua, F., Jones, J.P.G., Kasoar, T., MacFarlane, D., Martin, P.A., Mukherjee, N., Mumby, H.S., Payne, C., Petrovan, S.O., Rocha, R., Russell, K., Simmons, B.I., Wauchope, H.S., Worthington, T.A., Trevelyan, R., Green, R., Balmford, A., 2020. Insights from two decades of the Student Conference on Conservation Science. Biological Conservation 243, 108478. https://doi.org/10.1016/j.biocon.2020.108478

Gutzat, F., Dormann, C.F., 2020. Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice. Environmental Management 66, 435–449. https://doi.org/10.1007/s00267-020-01312-6

Haddaway, N.R., Bayliss, H.R., 2015. Shades of grey: Two forms of grey literature important for reviews in conservation. Biological Conservation 191, 827–829. https://doi.org/10.1016/j.biocon.2015.08.018

Haddaway, N.R., Westgate, M.J., 2019. Predicting the time needed for environmental systematic reviews and systematic maps. Conservation Biology 33, 434–443. https://doi.org/https://doi.org/10.1111/cobi.13231

Hahn, J., Todd, P., Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica 69, 201–209. https://www.jstor.org/stable/2692190

Hedges, L. V, Olkin, I., 1980. Vote-counting methods in research synthesis. Psychological Bulletin 88, 359–369. https://doi.org/10.1037/0033-2909.88.2.359

Hemming, V., Burgman, M.A., Hanea, A.M., McBride, M.F., Wintle, B.C., 2018. A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution 9, 169–180. https://doi.org/10.1111/2041-210X.12857

Iacona, G.D., Possingham, H.P., Bode, M., 2017. Waiting can be an optimal conservation strategy, even in a crisis discipline. Proceedings of the National Academy of Sciences 114, 10497 LP – 10502. https://doi.org/10.1073/pnas.1702111114

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge.

King, K.M., 2019. Can Google Translate be taught to translate literature? A case for humanists to collaborate in the future of machine translation. Translation Review 105, 76–92. https://doi.org/10.1080/07374836.2019.1673268

Kneale, D., Thomas, J., O'Mara-Eves, A., Wiggins, R., 2019. How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision making? Research Synthesis Methods 10, 44–56. https://doi.org/10.1002/jrsm.1320

Lawton, J.H., 1999. Are There General Laws in Ecology? Oikos 84, 177–192. https://doi.org/10.2307/3546712

Leclère, D., Obersteiner, M., Barrett, M., Butchart, S.H.M., Chaudhary, A., de Palma, A., DeClerck, F.A.J., di Marco, M., Doelman, J.C., Dürauer, M., Freeman, R., Harfoot, M., Hasegawa, T., Hellweg, S., Hilbers, J.P., Hill, S.L.L., Humpenöder, F., Jennings, N., Krisztin, T., Mace, G.M., Ohashi, H., Popp, A., Purvis, A., Schipper, A.M., Tabeau, A., Valin, H., van Meijl, H., van Zeist, W.-J., Visconti, P., Alkemade, R., Almond, R., Bunting, G., Burgess, N.D., Cornell, S.E., di Fulvio, F., Ferrier, S., Fritz, S., Fujimori, S., Grooten, M., Harwood, T., Havlík, P., Herrero, M., Hoskins, A.J., Jung, M., Kram, T., Lotze-Campen, H., Matsui, T., Meyer, C., Nel, D., Newbold, T., Schmidt-Traub, G., Stehfest, E., Strassburg, B.B.N., van Vuuren, D.P.,

Ware, C., Watson, J.E.M., Wu, W., Young, L., 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. Nature 585, 551–556. https://doi.org/10.1038/s41586-020-2705-y

Levins, R., 1968. Evolution in changing environments: some theoretical explorations. Princeton University Press, New Jersey.

Levins, R., 1966. The strategy of model building in population biology. American scientist 54, 421–431. https://www.jstor.org/stable/27836590

Light, R.J., Singer, J.D., Willett, J.B., 1990. By design: Planning research on higher education. Harvard University Press, Cambridge.

Maas, I.L., Nolte, S., Walter, O.B., Berger, T., Hautzinger, M., Hohagen, F., Lutz, W., Meyer, B., Schröder, J., Späth, C., Klein, J.P., Moritz, S., Rose, M., 2017. The regression discontinuity design showed to be a valid alternative to a randomized controlled trial for estimating treatment effects. Journal of Clinical Epidemiology 82, 94–102. https://doi.org/10.1016/j.jclinepi.2016.11.008

Mapstone, B.D., 1995. Scalable Decision Rules for Environmental Impact Studies: Effect Size, Type I, and Type II Errors. Ecological Applications 5, 401–410. https://doi.org/10.2307/1942031

Marshall, I.J., Johnson, B.T., Wang, Z., Rajasekaran, S., Wallace, B.C., 2020. Semi-Automated evidence synthesis in health psychology: current methods and future prospects. Health Psychology Review 14, 145–158. https://doi.org/10.1080/17437199.2020.1716198

Marshall, I.J., Kuiper, J., Wallace, B.C., 2015. Automating Risk of Bias Assessment for Clinical Trials. IEEE Journal of Biomedical and Health Informatics 19, 1406–1412. https://doi.org/10.1109/JBHI.2015.2431314

Marshall, I.J., Wallace, B.C., 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews 8, 163. https://doi.org/10.1186/s13643-019-1074-9

Metadataset, 2020. Metadataset [WWW Document]. URL www.metadataset.com (accessed 3.4.20).

Moher, D., Schulz, K.F., Altman, D.G., 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 357, 1191–1194. https://doi.org/10.1016/S0140-6736(00)04337-3

Moscoe, E., Bor, J., Bärnighausen, T., 2015. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. Journal of Clinical Epidemiology 68, 132–143. https://doi.org/https://doi.org/10.1016/j.jclinepi.2014.06.021

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.P.A., 2017. A manifesto for reproducible science. Nature Human Behaviour 1, 21. https://doi.org/10.1038/s41562-016-0021

Mupepele, A.-C., Walsh, J.C., Sutherland, W.J., Dormann, C.F., 2016. An evidence assessment tool for ecosystem services and conservation studies. Ecological Applications 26, 1295–1301. https://doi.org/10.1890/15-0595

Nakagawa, S., Dunn, A.G., Lagisz, M., Bannach-Brown, A., Grames, E.M., Sánchez-Tójar, A., O'Dea, R.E., Noble, D.W.A., Westgate, M.J., Arnold, P.A., Barrow, S., Bethel, A., Cooper, E., Foo, Y.Z., Geange, S.R., Hennessy, E., Mapanga, W., Mengersen, K., Munera, C., Page, M.J., Welch, V., Haddaway, N.R., 2020. A new ecosystem for evidence synthesis. Nature Ecology and Evolution 4, 498–501. https://doi.org/10.1038/s41559-020-1153-2

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716. https://doi.org/10.1126/science.aac4716

Parker, T., Fraser, H., Nakagawa, S., 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology 33, 747–750. https://doi.org/https://doi.org/10.1111/cobi.13342

Pototsky, P.C., Cresswell, W., 2020. Conservation research output in sub-Saharan Africa is increasing, but only in a few countries. Oryx 1–10. https://doi.org/DOI: 10.1017/S0030605320000046

Shackelford, G.E., Martin, P.A., Hood, A.S.C., Christie, A.P., Kulinskaya, E., Sutherland, W.J., 2021. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biology 19, 33. https://doi.org/10.1186/s12915-021-00974-w

Shapin, S., 1998. Placing the view from nowhere: historical and sociological problems in the location of science. Transactions of the Institute of British Geographers 23, 5–12. https://doi.org/10.1111/j.0020-2754.1998.00005.x

Sutherland, W.J., 2019. Kaizen conservation? Oryx 53, 397–398. https://doi.org/DOI: 10.1017/S0030605319000619

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19, 305–308. https://doi.org/https://doi.org/10.1016/j.tree.2004.03.018

Sutherland, W.J., Taylor, N.G., MacFarlane, D., Amano, T., Christie, A.P., Dicks, L. v, Lemasson, A.J., Littlewood, N.A., Martin, P.A., Ockendon, N., Petrovan, S.O., Robertson, R.J., Rocha, R., Shackelford, G.E., Smith, R.K., Tyler, E.H.M., Wordley, C.F.R., 2019. Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database. Biological Conservation 238, 108199. https://doi.org/10.1016/j.biocon.2019.108199

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, Steven, Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., Albrecht, L., Horsley, T., Platt, J., Armstrong, R., Nguyen, P.H., Plovnick, R., Arno, A., Ivers, N., Quinn, G., Au, A., Johnston, R., Rada, G., Bagg, M., Jones, A., Ravaud, P., Boden, C., Kahale, L., Richter, B., Boisvert, I., Keshavarz, H., Ryan, R., Brandt, L., Kolakowsky-Hayner, S.A., Salama, D., Brazinova, A., Nagraj, S.K., Salanti, G., Buchbinder, R., Lasserson, T., Santaguida, L., Champion, C., Lawrence, R., Santesso, N., Chandler, J., Les, Z., Schünemann, H.J., Charidimou, A., Leucht, S., Shemilt, I., Chou, R., Low, N., Sherifali, D., Churchill, R., Maas, A., Siemieniuk, R., Cnossen, M.C., MacLehose, H., Simmonds, M., Cossi, M.-J., Macleod, M., Skoetz, N., Counotte, M., Marshall, I., Soares-Weiser, K., Craigie, S., Marshall, R., Srikanth, V., Dahm, P., Martin, N., Sullivan, K., Danilkewich, A., Martínez García, L., Synnot, A., Danko, K., Mavergames, C., Taylor, M., Donoghue, E., Maxwell, L.J., Thayer, K., Dressler, C., McAuley, J., Thomas, J., Egan, C., McDonald, Steve, Tritton, R., Elliott, J., McKenzie, J., Tsafnat, G., Elliott, S.A., Meerpohl, J., Tugwell, P., Etxeandia, I., Merner, B., Turgeon, A., Featherstone, R., Mondello, S., Turner, T., Foxlee, R., Morley, R., van Valkenhoef, G., Garner, P., Munafo, M., Vandvik, P., Gerrity, M., Munn, Z., Wallace, B., Glasziou, P., Murano, M., Wallace, S.A., Green, S., Newman, K., Watts, C., Grimshaw, J., Nieuwlaat, R., Weeks, L., Gurusamy, K., Nikolakopoulou, A., Weigl, A., Haddaway, N., Noel-Storr, A., Wells, G., Hartling, L., O'Connor, A., Wiercioch, W., Hayden, J., Page, M., Wolfenden, L., Helfand, M., Pahwa, M., Yepes Nuñez, J.J., Higgins, J., Pardo, J.P., Yost, J., Hill, S., Pearson, L., 2017. Living systematic reviews: 2. Combining human and machine effort. Journal of Clinical Epidemiology 91, 31–37. https://doi.org/10.1016/j.jclinepi.2017.08.011

Tsafnat, G., Dunn, A., Glasziou, P., Coiera, E., 2013. The automation of systematic reviews. British Medical Journal 346, f139. https://doi.org/10.1136/bmj.f139

Vié, J.-C., Hilton-Taylor, C., Pollock, C., Ragle, J., Smart, J., Stuart, S.N., Tong, R., 2009. The IUCN Red List: a key conservation tool, in: Wildlife in a Changing World–An Analysis of the 2008 IUCN Red List of Threatened Species. IUCN, Gland, Switzerland.

Wallace, B.C., Dahabreh, I.J., Schmid, C.H., Lau, J., Trikalinos, T.A., 2014. Modernizing Evidence Synthesis for Evidence-Based Medicine, in: Greenes, R.A. (Second Edition. (Ed.), Clinical Decision Support. Elsevier, Oxford, pp. 339–361. https://doi.org/10.1016/B978-0-12-398476-0.00012-9

Wauchope, H.S., 2020. Working with large-scale population trend data in ecology and conservation: methods and applications. University of Cambridge. https://doi.org/10.17863/CAM.59354